Okay. Good morning, everyone. My recording has started. That means I have to start as well. It started at exactly ten. I know. Yeah, it starts automatically. So all the lectures, the recording start automatically at the scheduled time, so you have to be ready by then. Otherwise you have to start cutting the recording. And that's just a whole other. I know. I don't know why I have to do this. It's an odd part of my job. Okay. Thanks for the care. How are we all doing? Good. More or less good. It's starting to get a bit darker in Edinburgh now, so all of you who are not used to kind of dark and very wet and sad winters. Be prepared. It'll start soon. It's. Yeah, it's October now. So by November you'll all miss beautiful August and September weather, which was much drier than you would normally expect in Scotland. So I hope you're you're ready for that. Try to spend some time outside. It helps get a daylight lamp if you need to. Apparently they are really good, though. Yeah. Try to try to stay positive in in winter. Take a vitamin D supplement if you've spoken to your doctor because that can help as well. Good. Okay. I hope you're doing well. The first thing I wanted to do this week is talk briefly about your coursework. I know it seems a really, really long way off because it's like deadline is early December for this coursework, but I still wanted to start talking about it because I think if you now start thinking about it within your groups, it's much easier in the end to then make the deadline Well in time. You remember that you also have your principles of data analytics deadline next month, so you won't have much time for this coursework during that time, which is normal. So you have to manage both projects at the same time. If you start early, it's doable and you know that we have uploaded a brief I think a couple of weeks ago. I have also uploaded the data sets. Now there's two files. I don't know whether you've had a chance to have a look at them. You have the main data file, credit scoring data in TXT, and then you have your variable list, which is kind of give a data description file. If you had a look at that already, you will notice that it's not pretty. It does not look good. You're probably very unhappy with me. There is very odd missing data, unstructured data, weird variable descriptions which are incomplete and not really very easy to understand. That's on purpose. So if you think I made a mistake, I might have. I'm only human. I do make mistakes and creating data sets for you. But if you think this is very odd, it's most likely on purpose. So if you have, for example, a look at your variable list, you will see that there is these variable descriptions for each file. Let's open that in Excel. That makes it a bit easier. Hmm. Oh, really? There we go. Okay. So this is actually a quite typical data description file. You can see that we have our variable names in column B, you have a description and column C, and then you have the values that can take in column D and you will notice that some of them are just empty or it doesn't really tell you anything like this one. So this is all on purpose. If, for example, marital status encoding not informed and you know, it's some kind of ordinal variable, what that means is you have incomplete data descriptions for that variable. So what I want you to do in those cases is make a reasonable assumption and maybe a bit of critical thinking in that kind of context. You can, for example, make an assumption about what could these categories mean? Or you can say, I can't make a reasonable assumption about that. It's best to ignore that variable. Or you might say, let's just do an analysis and then they can, when they interpret the results, do the interpretation of them themselves. So there's many different ways you can go about that, but things like that are on purpose. So if you think this is mean that this is odd, that's just the way it is, you got to deal with it good. Now, if there are really errors in the data where you think, okay, this can't be it completely breaks my code or you can't even open the data set or anything like that, do let me know because there might be actual errors in the data set, but if it's just a bit inconvenient, then it's on purpose. Okay. Are there any initial questions about the data? Did anyone have a look at them and already went into despair about. No. Okay. A few nods. I see some people already had a look. What did you think? Good data. Bad data. It'll do. Yeah, it's not it's not easy, but I think it will kind of. Help you get a good understanding of what data might actually look like in the real world. So the whole purpose of this course is to give you. The tools necessary to work with real life data, which might not look pretty. The other thing is you might think, I can't start anything yet because we haven't done any kind of modelling in the lecture. What am I supposed to do? But I think if you think back to the first two lectures, each lecture actually covered something which is really important and which will become an important part of report. So I would argue you can absolutely already start working on that report writing that report. If you think, for example, back to the first lecture, we were talking about how to formulate research questions. What is a research question? What kind of modelling do I use for what kind of research question You can start describing the data. You can start describing the background the data might come from. What does the model even what's the model supposed to tell me before you even start making decisions about the model itself? You have to first decide what kind of model is suitable for this data or what kind of questions am I trying to answer, which is all coming from the brief. So you can already start thinking back to the first lecture and you can start thinking back to the second lecture where we talked about missing values and data imputation and outliers and all of that. As we've just noticed, there's quite a lot of that probably in the data. So you can already start with those preprocessing steps as well. So if you are quite keen on starting early, you can. That being said, this week we will actually start talking about a couple of quite simple predictive modelling steps that you typically undertake. So this week we will talk about linear regression, which is. Kind of the basis of understanding how predictive models might work. So the reason why we put linear regressions in the very beginning is, one, it's a natural connection back to what you've done in other lectures, what you've done in the past. And the other is it's really, really easy to explain how predictive models work with a linear regression because it's a very intuitive model to study. Yes. Before we start that, let's have a brief recap of last week. As I mentioned, we talked about data cleaning. We talked about preparation of data sets, preprocessing standardisation, variable transformation. You remember that was this whole one hot encoding integer encoding of variables that we were doing outlier detection. That was this idea of looking at the mean how many standard deviations away from the mean are my observations and missing values, how we can remove them or how we can impute them. Now this week we first have a brief, another look at descriptive statistics and visualisations why we do that. I will explain that in a minute and then we will cover linear regression in nice and easy steps and quite a lot of depth as well. This will double up with a lecture in principles of data Analytics, which you'll have in week five, where we also look at linear regression. But in that lecture we will look more into the assumptions and the statistical underlying principles of that method. Whereas here we look more at the applicability of that approach and what you have to remember when you actually use it in predictive settings more so than in descriptive ones. So we'll talk about simple linear regression, multiple linear regression and also any applications and considerations that you have to take into consideration. Yes, I doubled up on that. Okay. So why are we talking about descriptive statistics and visualisations? Again, we've done that in every single lecture. I think we've so far. The reason for that is describing your data as multiple really important reasons in any kind of data analysis, module, project, research, question, whatever you want to call it. Some of them we've already talked about. So we talked about how descriptive statistics, visualisation of descriptive statistics can be used to visualise and communicate your data. We've stressed how really, really important that is because it helps you to bridge the gap between your more technical knowledge and the technical understanding of data and modelling and that understanding of the person you're trying to communicate with. So if you're trying to communicate with someone, for example, in the management level, then you might want to think about what are they trying to answer and find out as briefly and as quickly as possible. They will not have. If you have to report to someone higher up, chances are they don't have time to read 20 pages of report, especially if there's a lot of numbers and just a lot of details on the model and all of these things. They're probably asking you, okay, what's the key takeaway that you want me to have from this report? And if you then point to a nice graph and explain it's this picture, they will be really happy with you. So try to think about your audience every time you do any analysis or any type of report writing. Who is your audience? Who are you trying to report to? Think back to this coursework assessment In this course, your audience will be a bank. So think about what is the bank actually interested in? Are they interested in a lot of technical details on the models? They might be, at least in that you've undertaken them with sufficient care, that you've tested the assumptions, etcetera, etcetera. They will be interested in that validity of the model. But think back to how you actually don't want to present the answer and the results to someone who might not have that interest in the statistical background. So that's visualisation communication. It's a huge topic, especially in a business school setting. The other thing is detecting outliers and the structure of data. So we were talking about outliers, which are basically arguing, for example, on two ends of a distribution. But you might also want to think about detecting outliers which are only accruing in a specific part of the data set. Now that specific part can be, for example, the tail end of a distribution, but it could also be that you're collecting data from different sources and some sources have a higher degree of missing. Values in them compared to others. That is really important to understand for you, because imagine that you're collecting data from two different sources and you think that both of these locations, for example, are equally important in understanding the overall structure of your data. Now, if one of these data sets, for example, has all missing values in one column, and then you're just deleting all the rows because you learned missing values, let's just delete it because that column is important. You're basically removing the data from one of these locations that you collected it from completely, which would then bias your model towards only responding to one of your locations. So understanding where your outliers are located, both in the sense that from which data they are stemming from, but also which variable they are cuing in or which region of the data they are queuing in, for example, or if you have a data set and you're looking at the income of a person and you notice that everyone in a higher income group is not reporting their education, then that will also bias your model. If you are interested in education as a variable, because then suddenly you're only able to analyse the education spread of lower income people in your sample. So think about where your outliers are occurring and then think about the reasons why they might be occurring there. We also have to understand the possible distribution from which the data stems. We've started talking about that in the principal's lecture. So thinking about your data distribution, where the data stemming from is very important for multiple reasons. Detecting outliers is one of them and making assumptions for your models is another. We'll also talk about how these can be used to evaluate model performance and to understand general trends and correlations in the data. So this is exactly the reason why we cover these this topic of visualisation and descriptive statistics again today. We will have a look at how we can use visualisation to evaluate model performance. And we will also look at how we can use visualisation to detect general trends and correlations in the data. For example, decide whether our data is linear. And if you think back to our main model that we were looking at today being linear regression, you can see how important it can be to decide whether the data is actually following a linear trend. Let's start with evaluating model performance. This is one of the topics that will come up again and again throughout the lecture series series. So we'll talk about how to evaluate a model probably briefly each time we discuss a new model. The reason for that is there's different approaches depending on how your model is actually performing their predictions, for example. And so depending on the approach, depending on the model, we can use very different approaches and very different measures. There's two which we will introduce today because they're commonly used with linear regression and you will also see them pop up again and again because they're really important concepts. The first is the mean squared error MSE and also the root square of that, the root mean squared error MSE. This is basically looking at an error rate. So we're looking at the deviation between any predicted and any observed values. And if I say predicted, I mean that in a kind of loose sense. So today, linear regression, the way we use it in the lecture will not be necessarily as a predictive model, as in we're trying to predict very specifically new records coming in with a new data set. But I mean by that that we will see how we derive parameters for that model which are then able to predict new data records coming in which are not necessarily the same. You will see the difference. You will see the difference. What I mean by prediction for the linear regression as we talked today and then prediction for new observations as we will talk next week when we talk about logistic regression. So just keep this in mind. Generally, you're just looking at deviation between predictive value and observed value, and we use that to evaluate our overall model performance. And overall is a key word here. The other thing is we look at R-squared, which uses the amount of explained variance to also evaluate the overall model performance. We look at this in a bit more detail a bit later, but first we will look at the mean squared error. So the MSE is, as I just said, an overall error and model performance measure. So you can always use that if the outcome of the model is numeric. That's very important to remember. The reason for that is we're basically calculating how far away is our observed value y from our predicted outcome for that value y hat. And then we sum up all of these deviations, all of these little errors, and we divide them by the number of observations that we have in total. And if you then take the root square of that. So if you take the root, if the MSE value that gives you the deviations also in the same units as were used in the sample. So you will see either the MSE or the square root of that in a lot of different software outputs as well. So it's quite an important value to to look at. Now you will hopefully immediately see how many times I said this is overall, this is overall. This is important to keep in mind because it does not tell you how well the model performs on specific parts of the data, for example. So we will we will talk about different types of errors throughout the lecture. So it's important to keep in mind that sometimes some types of errors are worse than others, depending on what you're trying to predict. If you're trying to predict, for example, illness in a patient and then you're missing someone, that's worse than predicting someone might have an illness and then later realising they do not. So there's different levels of error which are more acceptable than others. And in this case, we're just looking at all deviations in numeric sense, summed up over the whole data set. We don't care about whether there's some more errors in some areas than in others. We just send them all up. We divide them by the number of records. We look at the average error basically across the data set. So what does that actually look like? If we think about we might have some kind of scatterplot between, for example, our X and then some kind of value y. So let's look at a scatter plot like that. Beautiful. Let's say we predict some kind of linear relationship between X and Y, and then what we would do is we would sum up all these deviations, drop in a different colour. That would be cool. Hey, here we go. All these little deviations. Form the actual predicted value. So if you sum up all of these and divide them by the number of records that you have, that will then give you your mean squared error. The idea is that. The. You will already see what I was trying to say. If I said you only look at the whole overall data. So let's imagine, let's take let's take a different colour. Let's take purple. I like purple. So let's imagine that some of these, especially over here. Are much worse than down here. So you can see that our deviations up here are much larger than our deviations down here, which are really, really small. But if you take the overall measure of that, you wouldn't detect that. You wouldn't detect that in a higher space. The deviations are actually much larger. The only thing that you would detect is that the average deviation for your predictive values is some kind of one number, basically. So they can be used as measures for the overall model performance. But because we sum up all residuals over all data sets, we're not actually looking at the deviations in their specific spaces. So we don't know whether the model performs well in some areas and worse in others. And one way of actually detecting whether the model performs well in some spaces and worse than others is through visualising that. So we just had a look at a scatter. You could see how there was a bit of deviation towards the larger values, for example. And then if you visualise that you can see these deviations getting bigger, then you can actually say, Hey, my model is not well performing in these higher value areas of X. That is exactly this idea if if you have some kind of tunnel, basically. So if your data is spreading. So there's more variation being created in higher values. And you still have your one prediction. This would only be visible via visualisation. It would not be visible if you just calculated your error. And I think this is one of the problems that I sometimes have in machine learning as kind of a general area of research. I think people and using machine learning models tend to over. Emphasise the importance of error rates and accuracy as to numeric values. And just look at this one number and basically tell me the model is performing well overall and this is the singular number that proves that it's definitely the right model. So they might choose one model of another, for example, because it has a higher accuracy without thinking about whether it's actually performing well in all areas of the data set. And if all areas of the data set are equally important. Now, if you, for example, have one model which has an overall slightly worse error, slightly higher mean squared error, for example. So higher error rate, worse model, but it performs equally well over the whole data set. Then in some cases that would be the better model to choose, even though it has the worse error because it would actually emphasise both both these spaces, for example here. So if you have a model that is able to capture both these areas and these areas up here, one might argue that it's a better model for the data even though it has the worst error rate. So that's one of my pet peeves, is simply reporting a singular error rate or a singular R-squared value or accuracy value, whichever you're reporting, and not thinking about whether it's true for all records. So yeah, pet peeve. Please keep that in mind, maybe for your own modelling later. Oh, gosh. Oh, no. I'm talking. Lesson of the day. I can't breathe coffee even though I want to. Oh. Wouldn't it be wonderful if you could all breathe in caffeine? Oh, okay. Back to this. The other measure that I would like to talk about is R-squared. It's probably one of the most famous and important model performance indicators. You will see it pretty much everywhere. If you want any model in R, you will get some kind of R-squared. If you read any paper about doing any type of modelling, they will most likely report an R squared value. So it's pretty much everywhere. Being able to interpret it is a really, really important skill. So you will see the keyword. Again, it's an overall model performance indicator. So we don't look at specific areas of the data. We create an overall model performance indicator and you calculate it quite similarly to our error rate earlier. So we still have we still sum up the deviations of our observed outcomes y with each of our predictive outcomes for that specific value at, at I and we still square those, etcetera, etcetera. However, we now also divide them by the observed value minus the mean value of that data. And then we take one minus the whole fraction that we have here. And the another way of talking about that is actually the bottom one here. So you'll see that in that way or similar ways multiple times depending on how you actually spell the sum of squared errors, for example. And so R-squared can also be described as the sum of squared errors, which is also called the residual sum of squares and the sum of squares total, which is the total sum of squares. So you'll still also see it as SS and RSS, but it's the same concept and. Divided by each other. And then one minus that value. It's basically giving you the same thing. So the SSE we just said earlier are our squared errors. So all the deviations, for example, from our regression line and then our ssts are the sum of squares total. That basically means it's our deviations from the mean value. So how far does the spread the data spread around the mean value? We look at two different types of variation here. One, we look at variation around our predictions. So for example, our predicted line, how far away do our points spread? And the other is we look at the total variation. So how far do values spread around the mean if you. Nope, not yet. Apparently. I'm not doing that. Why am I not drawing that? How odd. Hmm. So you'll have to do with my hands describing concepts. Yes. So we have deviation around our, for example, linear predictor and we have overall variation, overall deviation from the mean. It is commonly reported in the form of a percentage of variation that is being explained by the model compared to the overall variation of the data. So if you see, for example, R-squared reported in in the literature, in many cases it's either reported at something like 0.7 or it's being reported something like 70%. And then the paper might say something like, the model explains 70% of the total variation in the data. So we have a total variation, variation around the mean. How much of that are we actually explaining that is the variation around you? For example, Linear Predictor. It suffers from the same problem as our MSE because it's overall evaluator. And furthermore, R-square also is less accurate for models with large numbers of predictors. So in that case it's usually recommended that you use the adjusted R-squared instead. The formula is here. It's still very similar, so you still have your normal R-squared calculation in here, but you then include the sample size n and the number of predictors in that formula, which then helps you to adjust If you, for example, have more predictors, they naturally explain more variation in the data. So the more data you actually bring in to your predictive modelling, the more of the variation is naturally explained, even though the difference might be very marginal and even though bringing in more predictors sometimes also leads to overfitting to the data and similar problems. So adjusted R squared basically takes into consideration that more predictors are not necessarily better for a model and therefore adjusts to the number of those. So if you have a large number of predictors, choose adjusted R-squared. It also doesn't hurt if you have a small number of predictors. So it's not like you can only use that for more than 30 variables or something like that. If you have just a basically large difference between your or not a large difference between your N and UK, then you could use that in in any way. So it's it's a safe choice to make whether you have a large number of predictors or not, but you have to use it if you do have a large number of predictors. Yeah, let's come to the second part of importance of visualisation. And we talked about how choosing a model depends highly on understanding your data. And that means multiple different things. But in this case, what I mean by that is the distribution of the data and more specifically the relationship of the data with the outcome that you want to predict. So we looked at the MSE earlier in terms of deviation from such a linear predictor, so deviations from our straight line basically. So how do we actually know that we should use a linear predictor? And sometimes the best way to decide that is simply through plotting your data. So plotting your data in the first step really helps you to understand whether you want to use a linear predictor or whether you have to use a nonlinear one. So if you, for example, have. A scatterplot which looks something like this. You remember that early on when we talked about clustering in the first lecture, actually. We talked about, hey, there's a bit of a linear trend visible in that data. So this is one of these cases where we can see a bit of a linear trend. Going, something like that. So that would be a case where I would say, okay, try a linear model. It might give you a good fit. In other cases, if your relationship looks something like this. Hmm hmm hmm hmm hmm hmm. I had a focus on art in school. So I was one of these terrible art students. Which you will now notice is beneficial for my job. Isn't that wonderful? So this would actually be a very non-linear relationship. In this case, you would look for a nonlinear prediction model, Something like a decision tree, for example, works usually pretty well. And then let's look at a third case. And I'm curious what you would make out of that. So let's let's remove a bit of this first. Give me a bit of space. Sit. It's. It's. It's. Okay. What would you do if you have a bit of a model like this? Huh? Let me make that a bit. Yeah, pretty good. What do you think? Yes. Like this? Yeah, you could definitely fit something like that. So you would fit, like, something non-linear. Just basically what you're saying. So something like this? Yeah. Very good. Anyone else? Another idea. Well, if you look at these separately, they do look pretty, pretty linear, don't they? So we have like a linear relationship here and we have a linear relationship here. You could actually do a piecewise linear model to that. So you could say you fit one linear model in this value space and then you have fit another linear model in this value space. So there's ways of using piecewise linear regressions, for example, which allow you to really fit quite closely to data which has linear relationships in different spaces of your value range. Why would you do that? Can anyone think of a good reason? Cheap. Yeah, it's really simple. It simplifies things, which means it's cheaper. So these kind of regressions, for example, then fitting regressions to different spaces of the data can be cheaper modelling wise, especially if you have a large data set. And another advantage is we will see that linear regressions can be interpreted really, really well. So if interpretation is a big thing, then thinking about how you can fit, for example, small linear regressions to your data might be helpful. There's also another thing that you can look up if you're interested, which is splines. That's really fun. I think I was told I was asked about time series analysis by one of your colleagues earlier. So if you have, for example, a Time series. Which goes something. I don't know. I don't know what time series look like, something like that. There's actually an approach where you try to fit splines to different areas of that data, so you would fit them, probably something like that and then piece wise them together. So that's another really interesting thing in connection to time series data. And it's a similar thought process because you're fitting these kind of segments together to replicate, in this case, how the time series goes and in the bottom case, how the data has linear shapes and different areas of your data space. So linearity linear models, it doesn't always have to be just this very simplified straight relationship that you're looking for. It can be kind of pieced together quite neatly. Brief mention of the bias variance. Trade-off In this context, we will go through that I think, next week in a bit more detail. But it can be quite important to start thinking about. So a high model bias can be thought of as the model oversimplifying a relationship. So what that basically means is to date, the model would not change even if you added more data to it. It's kind of very simplifying a relationship and a high model variance can be thought of as the model over complicating our relationship. So if you add small additional data points to that, the model will change significantly. In other ways to think about it are if I think about high model bias, I tell myself it's robust but inflexible. And if you think about high model variance, I think it's flexible, but it's sensitive. So the key for us is to find a model which basically gives us the best trade off between the two. So we want something that is fairly robust, so it doesn't change just if we add 1 or 2 more additional data points, but we still want it to fit fairly well to our data. How is that related to our pictures earlier? If we forget about the spline concepts and we just think that this is our regression line or our modelling line for some reason, and you can see how closely that line is following our data and how exactly this following it. Now, if you then make changes or have additional data data points later, then that line will change significantly because it's clings so closely to your data points. Whereas here if you add a couple more points here and there, it doesn't really change the line. The line just goes straight through it. So this kind of data, for example, here on the top left, that would be what we would describe as a quite model with a high model bias. So it's not significantly changing. If we add anything, it's not very kind of close to the data. I mean, it's fairly good. That's a really good model. I drew it, but it's still if we add more data to it, it's probably not changing much about the shape of it. So we will say it's not very flexible, but it's full bust, which can be useful depending on your context. So keep this term in mind bias variance Trade-off You will see that many, many times throughout the lecture series. It's also connected to the idea of overfitting models, for example. So if you overfit a predictive model, that means we fit it really, really closely to the data that we have, the training data. But then if we try to use it on unseen new data with small changes, then the model just breaks because it's too closely clinging to this idealised idea about what the data should look like. Okay, now I have a question for you. How can the shape of the variables be important for building a predictive model? So A, you need to know the shape of the dependent variable variable's distribution. B you need to know the shape of the independent variables distribution To build a model C, you need to need you need to know the shape of both to build a model or D, you need to understand the shape of the distribution because the need to to know the shape of the distribution depends on the choice of your algorithm. Which one do you think might be the most accurate? Are you saying B or D, D, d? Who would like to say answer the. Yeah. Okay. Who would think Answer be better. No. Any other opinions? A. See. No. So it's probably you. I think that's the most accurate. Yes. It's probably between C and D, So if you do need to understand the shape of your data for that model, then you need to know the shape of both. If you choose a model where you don't think the shape of the variable distribution matters at all, then you probably closer to answer D. So I mentioned earlier decision trees, for example. They are fairly good example for where I would say the distribution of the data does not matter that much. We don't have to check for linearity for example. However, we also learned that the variable distribution also tells us things about outliers. So you might argue that if you use the distribution of the variables to detect outliers, for example, then it would still matter even for a decision tree. They're not super sensitive to outliers, but they can be especially if they don't try to split on that specific variable with with the outliers. So in that case, you would need to know the variable distribution, even if it's not for the assumptions of the model itself. Okay. It's a 10:45. Let's take our ten minute break now because then I can put all the regression stuff into the second half, which kind of makes sense structurally. So I'll see you in ten. Okay, We want. Let's come back in. Please take a seat. Take a seat. Take a seat. Sit down. Sit down. Just. Just. Okay. I had the realisation that I did not stop recording, which means I have to go back into the recording later and cut out the break, which is very tiring. I'm a video editor as well. Honestly, if you work at a university, you have all the jobs at the same time. It's wild. I have to do research and I have to do a lot of admin tasks and I have to teach and apparently I have to be a video editor as well. So yes, that's my complaint from the break. The other thing is there was a question coming up a couple of times, which is about this descriptive and statistics visualisation stuff. If you have a lot of variables, do you honestly go in and check for each of these variables? What's up between them and the predictors and you do your little plots and everything? The answer is it depends. It really depends on your time. It depends on the model you're going to use. It depends on how important that relationship is for the assumptions that your model makes. And it depends on the number of variables that you have. If you're building a linear regression with 200 variables, then you're not going to check the relationship between each of them with the predictor unless you have the time to do so. The reason why you should is that the model itself by structure, as you will see, assumes a linear relationship between the predictor and your outcome variable. So if that relationship does not exist because it's linear, then fitting a linear regression is not going to give you the best possible predictions. That being said, if we are being realistic, you're not going to check that assumption for each variable necessarily unless you have the time to do so. So sometimes you might think, okay, let's try a linear regression because it's very quick. Let's also run a random forest, for example, on the same data. Let's compare the outcome of both of these. Oh look, the random forest is performing so much better. Might the reason be that the linearity is not there between the predicting variable and the outcome? Possibly. And then you can go in and check whether that's actually the problem. Now, I'm telling you all this and I can feel my statistician colleagues glare at me angrily because technically linear regression has a set of assumptions and if you violate these assumptions, then you cannot use a linear regression model because it's not going to give you a meaningful output. Now, the linear relationship between dependent and independent variable is not one of the strict assumption that have to be fulfilled in order to be allowed to use a linear regression. There are such assumptions though. We'll see them later during these slides. So you do technically, and I'm telling you that because I'm your teacher and I have to teach you stuff, you do technically have to check the statistical validity of each of your assumptions that we'll see here for your data before you use a linear regression. And now I'm taking a step back, and I'm telling you, as someone who's gone through it in the real world, you should test the validity of your assumptions. Because if you don't, then you start getting into this kind of let's just use any model without thinking about the assumptions, and let's just use the model and check the result. And if you start that kind of. I don't know how to describe it. You get into this trap of doing that for a regression and then you do it for other models as well. Then you can't really use your results in the end because they are not statistically valid. So then you're producing results which are not holding up to scrutiny. So yes, you should check the assumptions for each of your variables, even if there's a lot. But you don't necessarily have to check, for example, the linear relationship between dependent and independent for each of them. If you have 500 variables or something like that. So what does that mean for your course? If you use a linear regression, it would be prudent to follow the steps of assumptions that we're laying out here if they don't all fit into your text because there's a limited word count and I'm not going to read anything that is longer than the word count. Then you have an appendix where you can explain you have checked those assumptions. Here are the plots to prove it. So then you can actually prove to me that, you know, the assumptions are important. You checked them because this is a university and you have to go through the four plus steps of doing your analysis. Okay. I hope that more or less answers the question. Something I do with descriptive statistics, if I have a lot of data, is I show I demonstrate that I did the analysis by showing, for example, a part of it in the main text. So for example, I might choose 1 or 2 interesting variables, show a scatterplot to demonstrate that yes, I have checked the distribution of the variables. He is proof. And then I put all the other plots into the appendix. And if anyone wants to check whether I've actually done it, I have. Okay. All of that being said, let's go through the theory of linear regression and. Let's start with a simple linear regression here. So the way I've split these lectures is we're doing linear regression this week and logistic regression next week, and we're doing simple and multiple versions of both each time. Originally, the lecture structure actually had us do simple linear and logistic regression in one, and then multiple linear and logistic regression in the next lecture. So let me know whether the structure works better because I think it makes more sense to focus on just linear regression today and then just logistic regression next week. But I'm open to to input whether it might make more sense to mix those two up. Linear regression. I hope you're all roughly still familiar with this kind of equation. What we're trying to do is we're trying to fit a line in a two dimensional space, which is created by a singular, independent and a singular dependent variable. And we're trying to fit a line in such a way that it's closest to all data points. So the equation can therefore be put very simply at we're trying to predict Y, which is the AI dependent independent value, which we're trying to estimate with our intercept beta zero. And we have our x i so that the ith observation of our independent data, we have an actual parameter here which is describing how strong that relationship or that input of X is on Y. And then we have an error term because nobody is perfect and our model is not perfect as well. So there will be unexplained variation and error in that as I said, the goal of fitting a linear regression is we're trying to estimate these two parameters that we have beta zero, our intercept and better one, our parameter describing the relationship to X in such a way that we're minimising the distance between the data points and the straight line. So you will remember what that looked like earlier, where we had our two dimensional space, we had our scatterplot of data and we're fitting a line and we're trying to minimise the deviations and we're minimising the error from each of these points to that line. So that might mean we have to kind of shift the line a little bit. And if we, if we change the angle of the line, that is changing our beta one, and if we're shifting the line up and down, that is our intercept beta zero. So we can then write this estimated regression line as our beta, as our y hat, which is our estimated values of Y, and we are estimating them using the relationship between better beta zero hat, our estimated intercept and our beta one hat. Our estimated predict our estimated parameter for our predicting variable. Wow. We will also recognise this term down here. So this is our deviation between our real values Y and our estimated values y hat. And we also call these our residuals. So they are basically our error. Anything that we're not able to explain in our relationship and our kind of perfect relationship up here that we know is true, but there's always a bit of an error in there. And whatever that is, is the deviation between our estimate where we estimated values from our true values. Okay. So, yes, you will remember that. Let's still draw it because it's an important concept and I enjoy drawing now that I found out how to do it. I'm still so proud of myself regarding that. Okay, so. We've already mentioned. I have a bit of a scatter plot here, for example. We can see it's a linear relationship, so let's draw our regression line through that. Then this is our intercept. So this is our beta zero. It basically describes how high or how low the value would be for an X of zero. So if X is zero, how high would our value for. Wow, that's terrible. Okay. For why still be that's our better zero. It's basically the external value of why that is not impacted by X. Our default and then our slope of X is then described by our better one in this case. So if you think about it, it could be a very flat or very what is the English word for that? That's slope, the opposite of flat. The inclined. Yeah. So strongly inclined. Do we have a native speaker in here? All right. Yes. So very inclined slope. That. That's what I mean. So that would be a high beta, basically. So the question is how strong is the relationship or the impact of our X on our Y? Now, this is for your simple linear regression and we will see later with multiple linear regression, we will just expand that into more dimensions. That's probably where I will reach my limits of being able to draw. How do we actually find this best fit line? We've just decided that we have to estimate our beta zero and our beta one. You cannot estimate your residuals. That's just whatever is kind of left over at the end. So one of the most common approaches for estimating your parameters is using the least squares approach. So the question is among all the possible lines that would fit somewhere through the data, for example, through your points of the scatterplot, find the one that actually minimises the sum of squared errors. That's very logical. We thought earlier, okay, the deviation of each of these points from the line. If we sum all of those up, that's your sum of squared errors. So let's minimise that error that we are finding. You will also remember that we were talking about the problem with just looking at the sum of squared errors, namely finding a line which fits through the data as a whole the best. So we are not thinking about any deviations and specific areas of the data. We're just trying to fit the best line overall. So what does that look like in mathy terms? We are looking at the sum of the squared errors E, which should probably be epsilon. Okay. Which is the as we know, the deviations of each true value from each predicted value sum squared, that sum that all up over all data points. And we also know that we've just decided that our y hat is actually best described by this linear expression here, which is better hat zero plus beta hat one times x. So this is the term that we're trying to minimise. So this is our, this is our, our squared errors and with the least squared errors, obviously we're trying to minimise that error. So what are the values of beta zero and beta one which minimise this expression here? Now I actually show you how exactly you would calculate that by hand. So hold on tight. If we take the first order conditions of our problem, you remember from school times how to find a minimum of something to take the first order condition, you set it equal to zero, and then you solve it for your variables. So in this case, if we solve this beautiful expression up here, we have one which we're solving for beta zero and we have one which we're solving for better one. If you do all the steps on paper, it gives us the condition that beta zero hat should be the mean values of your observations. Minus beta one times x mean. So these little bars always denote your mean value. And your second condition tells us that beta one hat can best be described as your deviations of each of the values from its mean times. The deviation of each of your outcome variables from their mean and divided the whole thing again by your deviation of each of these values from the mean squared summed it all up. And if you fulfil these two conditions then you find the least squared estimates for better zero and better one. So if you ever have to solve a linear regression by hand, I don't know whether you I don't think you ever no need to. I've never done that by hand. This is the way to actually do it. So there is a way of solving that by hand. There's also technically a way to solve multiple linear regressions by hand, but they get quite complicated. So at some point you're not going to do that. But there is a way. So if anyone ever tells you you're sitting on a lonely island with no computer for some reason, but you want to calculate the relationship between a time of sundown and the growth of coconuts or something like that, and you could actually do that by hand. I have a number of example here for you as well, which I think will illustrate that quite nicely. So this is what this would actually look like if on the right hand side you can see a sample data set, which we have. For example, this is a box of a sales, something like it could be the sales of movie one and the impact that has on, I don't know, the profitability of the cinema or something like that. So let's say X is our movie sales and then Y is our profitability of the company. You will see they are lower. I've actually. Okay, now here, story time. I've actually worked in a cinema as I was while I was a student because I was really into movies. So I was working in a, in a really small town cinema, one of these where you have like we had two locations actually. It was a really small Bavarian town and we had two locations and I had to run between the two locations back and forth whenever we only had limited staff and then they ran out of, I don't know, chocolate bars, so I had to carry chocolate bars through the city. Best summer job ever. I got free tickets to the movies though. So that's that. Okay. So I'm very that's why I say box office sales profitability of the cinema is actually quite low. Cinemas don't make a lot of movies because cinemas have to pay to actually kind of rent the movies they are showing. So a lot of the money that you're paying for the ticket goes into being able to rent the movie. So that's also why smaller cinemas only show a limited number of blockbusters. They are really expensive to show, so they make money mostly through chocolate bars, which I was going back and forth. So please buy chocolate bars when you go to a small cinema. That's why these numbers are small. That's what I wanted to say. Okay, So how do we then actually describe the relationship between ticket sales and profitability of your cinema? We will start by calculating your mean values so you can see X bar is the mean value of X, So the ticket sales y bar would be the mean value of your profitability of the cinema. And then you would calculate if you remember back all of these expressions here, we just calculate them separately. It's the easiest way to actually be able to sum them up. So we create a table and we we calculate each of our numerators and denominators. So for example, the deviation of each value from the mean for each of these values 29, 49, 89, etcetera, we take that value minus its mean. We also take that value minus its mean squared for each of our y's that value minus the mean. And then in the end we take the product of those two. Of these t of these two terms. So we create a table for all values. It's not a lot, it's only five each. So that's relatively doable. And then we can actually calculate our beta hat one which comes out to 0.1481 and our beta zero, which then comes out to 14.179. Now if we then put those into one singular equation, it would look something like beta sorry, y is equal to 14.179 plus 0.481 times x plus whatever error term there is left over. So you would actually be able to calculate the profitability of the cinema as an equation between the 14. That's your intercept. So your intercepts basically external. The cinema is just making a profit as it is for some reason. And then depending on how high your ticket sales go, there's another factor coming in which increases it increases that profit. You better could also be negative. In that case, with each sold ticket to the cinema would make a loss. Okay, Now let's have a look at another example for that. So let's imagine we collect a sample of 209 salaries and company company profitability data from from the US. And if we have our formula earlier so we can calculate our beta zero and our beta one, we find this regression model here. The sell the estimated salary is calculated by 963 plus 18.5 times whatever is the profitability number of your. Of that company. So the question is, what would be the predicted salary of an executive? B whose company has a profitability of 0%. 963.2. Exactly. So that's basically the external value. So there's always that value of 963. So now we run our model in Excel or or in R or Python or whatever, and we learn that our R-squared of this model is 0.0132. Who can tell me What does this value mean? Please raise your hand. It's so difficult to hear two, three directions. I know. It's like being back in school. Yes. So we only 1.32% of the variation in. Exactly. So only one. So if we want to express that in percentages, we move the comma 1.3% of the variation in the model or in a salary is explained by this model. So there's a lot going on that is not part of either profitability or whatever is in this external value here. So what does this value not tell us? But they're just picked up on the side of a tie, you know? Is there any other variables left? Yeah, It's basically not telling us anything about what is not in the model. Right? It's basically telling us, okay, if we just have salary, we can explain 1.3% of that kind of. If we just have celery in profitability, we can explain 1.3% of the relationship between the two. We can't say anything about what other factors there might be. We also can't say anything about whether for some companies that value is higher or lower. So that was this idea. If we just look at the overall fit. We don't know whether, for example, that model works really well for companies on the East Coast, but not at all at the West Coast for some reason. So this is just the overall US model. We also can't say anything about whether this model would fit well for other countries. So we have only the US data to create this model. It might work better or worse or not at all for other country data. And that depends on whether the relationship is similar. Okay now. Let's say we don't have to calculate our linear regressions by hand. Let's say we actually use some kind of software. I'm pretty sure this is R, so this would be an output that you get from running a linear regression. R It looks the same in Python. You will see doing your tutorials tomorrow. Yes, you have a computer tomorrow and the computer tomorrow that the output looks very similar. So it will give us. The coefficients. The constant here refers to your beta zero and the x coefficient refers to your beta one. So this would, for example, be a model where we just explain something in terms of 14.17 plus 0.14 times x. It also gives us the standard error of those coefficients, which is really interesting because it tells us how kind of variable or how spread this coefficient is across the data. So how close or how how close are we actually to that one value is the variation across the data set in that one, in that one value or is the coefficient we can see, for example here the x coefficient beta one has a really small standard error, so there's very little variation in that. It's always kind of the same across the data set. So, yes, remember that these are estimates. We don't know the true value. And the question is whether we would get another value, a different value in another set of observations. What was the other? Yes, the accuracy of the coefficients. This is quite important. Back here we have a confidence interval being reported typically. So that basically means if we repeat that experiment on different sets of data, how likely is it that we are still around that similar coefficient? So you can see here, for example, our coefficient of estimated beta zero is 14 and we have 95% confidence that it will be between -6 and 35. So this is basically giving you a range within which you can expect that value to be even if you repeat that experiment with different data. And we also have something really interesting. We get information on hypothesis test. So hypothesis tests, you will cover them in depth tomorrow in principles of data analytics. And yes, they are being used in this lecture as well, mostly to look for the significance of your coefficients. So in this case, we are interested in whether our relationship between our predictor and our response is actually significant. We have the null hypothesis. So our thought process is our default is basically our relationship. The relationship is not significant. So our better one, for example, would be zero, and then our alternative hypothesis would be better. One is statistically significantly different from zero. And that's what we are testing for. Though you will hopefully remember something about hypothesis tests. This one is using a T test because it's a sample, etcetera, etcetera. So we are looking for the T statistic. That's your T column and we're looking for the P value and if you will all hopefully also remember that you want your T statistic to be high and you want your P value to be low because the T statistic is basically telling us whether our data is consistent with our alternative hypothesis, which means the D predictor or the beta is significant or basically not necessarily large, but significant in its impact. So you want the T statistic to be very large for that and your P value is telling you something about whether you should reject the null hypothesis in favour of your alternative hypothesis. So that's it's saying here the P value is smaller than 0.05. That's a really it's a rule of thumb. You can choose any p value that you like. Students often ask me, do I have to always use 0.05? No, you can choose any. Honestly, theoretically you could choose P values three. Of course, that doesn't make much sense that no one will actually do that. But you could. The P value is your chosen threshold underneath which you say I think this is enough evidence. You make the judgement, this is enough evidence to reject my null hypothesis. And this is just an indicator that is basically giving yourself a cut-off value that you have to follow can be 0.05 can be 0.01 could be 0.1. These three are the most common values that you will see. So in this case, we would say, for example, we can see our P value here is 0.08. So if we choose a threshold of 0.1, we would say the coefficient is significant at a 10% significance level. You can also see it's not smaller than 0.05 and I do not ever want you to say something like it's almost significant or it's close to significant. I will burn the paper down. No, I won't do that. That's that's too much. But I will be very angry if I ever see that there is no such thing as close to significant or almost significant. There just isn't. We just said it's your threshold. You decide. You say, this is my level and it's either underneath that level or it's not. It's yes or no. There's no in between. I've seen that. I see that in papers sometimes. So I think this is actually a big problem in research. You will you will hear my colleague hopefully rant about that tomorrow that people actually do stay close to significance when there's no such thing. It's significant or not binary. Okay. We've talked about this simple linear regression. Now let's have a brief look at our multiple linear regression, very similar concept. We're just extending it to include more than one predictor. So we take our simple linear regression and now we including not one predictor X, we including K of them and K can be any number that you choose. So depending on your data set size and we still have the same components, we still have our intercept, we still have our error term epsilon at the very, at the very end. So we still have the same logic behind that. We still look for the estimates of your coefficients, that's all your betas now. So all your better hats and we try to make a prediction for any instance X and its value y hat. So same idea. We just expanding it to include more. Now you can write that in matrix notation as well, which can be very useful if you think about that in data set terms and data frame terms. So if you think about having, for example, your Excel spreadsheet with your with all your data, then with all that would be all your X values. So you have, for example, N observations 200 people answered your survey and you ask them ten questions. Then you have K is equal to ten and is equal to what did you say? 100? 200? I don't know. Your number of applications. I just tell you random numbers, then forget them. I should write them down. That's terrible. Okay, we have our vector of betas now, so we have one better for each of our predictors, for each of our variables. K And we also have an error term for each of our observations. So there will be vectors of vectors of errors, vectors of coefficients and your matrix, which is basically your data frame. We also follow the same approach to finding our betters. So we still use our least squares approach and we just slightly change our calculation to include now all of our different X and all of our different betters. So we have K betters that we actually trying to calculate. We still can compute the first order condition to minimise our our our sum of squares and solve that for betters. And now we can see that our beta hats, remember this is a vector can actually be calculated like this and now you have your transformed matrix X times your matrix X. Take the inverse of that. ET cetera. ET cetera. And this will then give you your beta hat. That is my my colleague thought this is a really good idea to remember. I thought this was really cute. So remember, better hats are blue because the best linear unbiased estimator that this is basically it's your best fit for a linear model which is unbiased and is your estimator for that. I don't know whether you find it helpful. I find it a really nice tidbit. So just remember that if you use your least squares methods, you're trying to find the best set of batters for that data set to predict your set of Y's. Okay. Now we still have a bit of time left. And what I want you to do now is wake up a little bit because it's been a long day, so I would like you to think about the following based on what we've discussed today in the lecture. But you can also use your own experience, because I know a lot of you do have a background in analytics, so you will have used linear regressions in the past. Think about a couple of advantages of linear regression models that you can think of, think about, of some disadvantages of those. You can start by thinking on your own, discuss them with your neighbour or little group of 3 or 4, I don't really mind. And then in like five, ten minutes, let's say, let's say seven minutes time we will share with the class, collect a few of those advantages and disadvantages. The reason why I'm telling you that or why I'm having you do this is in the very beginning of this lecture series I emphasised The important thing for me is to critically think about what model is suitable, in which case, what are the advantages and disadvantages of each of these models. So we will do a similar discussion after introducing each model. And then at the end I want you to have a look at your notes. I would recommend that you that you write notes. Why will discuss that and why your colleagues say things because that helps you to develop your critical thinking skills in order to be able to critically discuss the models and compare them. Okay, so take a couple of minutes and then we'll share with the class. Also share with the class always makes me makes me sound like a middle school teacher. So sorry about that. Let's start recording again. Who would like to start? You can start. Advantages. Disadvantages or just general comments that you're not entirely sure fit into one of these categories. And we will go through the room roughly so I can hear from everyone. Let's start on this side, the front one. Have you discussed? You know. So one of the points that we discussed as an advantage is it is very interpretable to anyone. So if I want to code it, it is very understandable as well as if I create a model and show it to other stakeholders. It is very easy for them to understand that what we have actually done, but we took that as so basically we took that as it is simple to understand, but it is sometimes very simple to apply it to complex problems and maybe our solutions will not be as required so that we took it as a disadvantage. It is sometimes gives us very linear solutions to maybe nonlinear problems. So basically it is simple, but it can be simplistic. And not giving us a proper solution to the required problem. Basically. Yeah. So you will remember that in the first lecture when we were talking about which kind of models we would cover throughout the lecture series, we're basically increasing our complexity as we go along. So we'll start with linear regression, simple linear regression studies, which are very simple, as I said, easy to understand, easy to interpret, easy to explain to someone because it's a simple linear relationship. I mean, we're talking about this, for example, this example between sales and profitability or some there's 1 or 2 inputs or a couple of inputs and you have one output. You can interpret your your factors, your parameters, your betas. You can very simply explain if you increase this by ten units, this is the exact numeric effect that it's going to have. But you're right, it can be quite simplistic. It's a linear model. We will see much more complex nonlinear models later. Neural networks, for example, are extremely complex. They're basically on the opposite end of this complex complexity scale. So great for highly complex relationships, but terrible at actually being able to interpret and explain what's going on. So it's it's a bit of a mix of both. You have to you have to balance. Very good. In the back part of that group. Wonderful. In the back, on the side place. What have you discussed can be advantage, disadvantage. General comment because how can I say can do by. Putting a separate point into one line and it can show you very, very simple or very trick of view of the. Plot and you can see the tendencies. So it's kind of a visualisation advantage that you can visualise the result of that as well. Yeah, I think that's a good that's a really good advantage. It also kind of part of this interpretable interpretability thing because you can you can show relatively easily the relationships. You also we just discussed this kind of piecewise linear regression thing as well. So even if the relationship is generally linear but might be kind of divided into steps, then you can still use a linear approach to that. If you, for example, divide the data space into into different parts of it and then you can fit lines to different parts of the data and the interpretation of that, I think I was asked about that from one of your colleagues. If you, for example, have one linear relationship in a specific area of X and then a different linear relationship in a different area of x, and an example of that might be, for example, the relationship between age and income, where you have a general increase in income as your, for example, your sample of customers grow older and then you suddenly have a higher step. So there's something happening at this point in age where suddenly we have a stock increase and then again this general linear relationship. And you could, for example, explain that by, oh, people around this age are being promoted or something similar. So you can try and find interpretation for these kind of these steps in in your piecewise linear regression as well. Yes, very good. Do we have another advantage or disadvantage from the back or from the side where you kind of working together? Vastly? Yes. This is like this model is really robust when it so it's like simple. It's like used in almost every sort of financial application, be it like pattern modelling. We like accounting bit like predicting risks or like sort of equity in CSR and stuff. Like it's like really robust. I mean it finds itself almost being in every sector though. It's like very simple. Robustness. Like that? Yeah. This is a big one. And I think this also comes back to this variance bias trade off thing that we were discussing earlier in the lecture. So a simple model can be very robust, very easily to apply in a lot of different contexts, as you call it. As White pointed out, that you can there's a reason why we we mentioned this first. It's kind of the core of a lot of modelling exercises. So in many cases you will have that as a baseline model or you make a comparison to this as a basic first approach. And then in many cases it's all you need. Sometimes you don't need a complex model. Sometimes an easy model can be completely sufficient for explaining relationship and it can be used in a lot of different contexts. Okay, Advantages. Disadvantages. Somewhere in the back and forth in the middle. Yes. One of the things we said is it's only for numeric variables. Variables. You can't use anything else. A categorical. As an outcome. Disadvantaged? Yes. As an output or as a You can't. It's only numbers. Oh. Yes, for the output, yes. So you can only predict a numeric output. Numeric output. I mean, you can use categorical variables as predictors if you, for example, use a dummy encoding or Simula that we use. Yes. Yes. So the output has to be numeric. We will see next week when we talk about logistic regression that there we have a categorical output as a variable. So we will see that there are ways around that. But for the simple, the simple linear regression that we discussed today, it's numeric outcome. Exactly. Yeah. Mentioned how you have to assume that the independent variables or the predictors are not affecting each other, that they're separate. Yeah. So you have the assumption of independent predictors so there's no collinearity between them. For example, if you have perfect collinearity between your predictors, the model will actually completely break. So I had this happen to a lot of students where they try to run a linear regression and for example, they created all of the dummy variables and they did not remove one of them. So they are dummy variables were perfectly correlated. And I just told them, No, I'm not doing this. And they came to me and were like, why is I not working? What's the problem? The problem is if you have perfectly correlated variables, it completely breaks. And if you have strongly correlated variables, then your output is not really reliable. So you can't really trust the outcome of your linear regression if there's correlation between the predictors. So one of the important first steps I would recommend doing before you use a linear regression is create a correlation matrix. Look at how correlated all of your independent, all of your yeah, all of your independent variables actually are. And I think you will do that in yes, you will do it in the lab or you can do that in the lab. So hint hint. My solution for the lab tomorrow includes creating a correlation matrix so you can have a look at that as well. It's really useful and it's also a nice kind of first analysis step where you look at the relationship not only between your independent variable on the dependent, but you look at the relationships within your data set, and then there's a ways of overcoming highly correlated factors. So for example, you could use a dimension with dimension reduction technique where we then create factors or principal components.