Overview of regression analysis and linear modeling

Regression analysis is a technique for modeling data. In general, the goal of any statistical approach to modeling data is to take a sample of data that represents a population, then to use that sample to estimate some facet of the population. In the case of regression analysis, the focus is on the "conditional mean value" (defined below) of a single dependent variable y corresponding to a given set of predictor variables $x_1, \ldots, x_p$. We will start immediately with an example:

Example
Suppose we are interested in blood pressure in a population of people (say, all adults in the United States). If we use regression analysis to pursue this interest, the dependent variable in the regression analysis would be blood pressure. The predictor variables would be selected according to the project goals, but might include variables that are presumed to predict blood pressure, such as age and body mass index (BMI). In our population of interest, many people may have nearly identical age and BMI (i.e. there are many 35 year old people with BMI equal to 25). Among these people, blood pressures will vary. That is, just because we know a person's age and BMI, this does not mean that we will know their blood pressure with certainty. However, there is a mean (average) blood pressure for all people with a given age and BMI. This conditional mean is the main focus of regression analysis. The conditional mean is a function, because it depends on the predictor variables. In this example, the conditional mean is a function of two variables – age and BMI.

Regression analysis is a set of techniques for taking a data set that represents in some way a population, then using these data to estimate a conditional mean for that population. There are many forms of regression analysis, and new approaches are being invented all the time. Here we will focus on what is perhaps the oldest but still most widely-used method of regression analysis, known as linear least squares. Linear least squares represents the conditional mean function as a linear function of the predictor variables. Suppose our dependent variable is y, and the predictors are x_1 and x_2. Then in linear least squares, we model the conditional mean function as

$$E[y|x_1, x_2] = b_0 + b_1 x_1 + b_2 x_2.$$

Here, the notation $E[y|x_1, x_2]$ is read "the conditional mean of y given x_1 and x_2" – "E" stands for "expectation" which is a synonym for "mean" (also equivalent to "average value"). The symbols b_0, b_1, and b_2 here are called the regression parameters. The regression parameters are numeric values that define the conditional mean function. They are not known, and must be estimated from the data.

Interpreting the regression parameters

The parameter b_0 in a linear model is called the intercept, it is special because it is not multiplied by any predictor variable. The other parameters are often called slopes. The slopes tell us how much the average value of y differs when comparing individuals in the population that differ by one unit for a particular predictor variable, but are the same in terms of all other predictor variables. For example, suppose we compare the average blood pressure (y) for someone whose age (x_1) is equal to 35, and whose BMI (x_2) is equal to 26, to the average blood pressure for someone whose age is 35 and whose BMI is 25. The conditional mean blood pressure for the first individual is $b_0 + b_1 \cdot 35 + b_2 \cdot 26$, and the conditional mean blood pressure for the second person is $b_0 + b_1 \cdot 35 + b_2 \cdot 25$. The difference between these two conditional means is b_2.

In many cases the intercept parameter is not very interpretable. The intercept is always equal to the conditional mean of the dependent variable when all of the predictor variables are equal to zero. However for some variables, the value zero may fall outside the range of interest, e.g. in the blood pressure example we are not interested in the blood pressure for someone with age and BMI equal to zero (which is an impossible value). Although the intercept is often not interpretable, including it in the model helps to give a more realistic representation of the conditional mean for many populations, so nearly all linear models include an intercept.

There is one special setting where the intercept is interpretable however – that is when all of the predictor variables have been centered to have mean zero. In this case, the intercept is the conditional mean of the dependent variable when all of the predictor variables are at their mean value, e.g. in the blood pressure example it would be the average blood pressure for someone who has average age and average height.

Terminology and notation

Regression analysis is a big topic that has been under development for more than a hundred years. Over this time, a variety of different terms and notational conventions have arisen. We will try to use common and standard terminology here. For reference we have included a table below that provides some common alternative terms for the key language used in this document.

In addition to there being different terminological conventions, there are also different notational conventions. Above we expressed the regression function in terms of the conditional mean $E[y|x_1, x_2, \ldots]$. This is one common way to express the key object of interest in a regression analysis. An alternative way to express a regression model is in "generative form":

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \epsilon$$

In the expression above, ϵ represents unexplained variation (also called "random variation", "error", or "noise"). The presence of the error term ϵ is necessary since the data we observe in the real world will never follow a linear relationship exactly. These "error terms" are used to represent the part of the data that cannot be explained using the predictor variables.

Term used here	**Common alternative terms**
Dependent variable	Response variable, outcome variable
Predictor variable	Independent variable, covariate, regressor
Mean	Expectation, average
Regression parameter	Slope, coefficient, effect
Unexplained variation	Error, noise
Conditional mean function	Regression function
Constant variance	Homoscedasticity
Nonconstant variance	Heteroscedasticity
Linear model	Linear regression model, linear statistical model
Fitted values	Predicted values
R-squared	Proportion of explained variance, coefficient of determination

Linearity

The role of linearity in a regression analysis is more subtle than it appears on

the surface. The conditional mean model used in a linear least square analysis is linear in two senses: (i) it is linear in the predictor variables, and (ii) it is linear in the regression parameters. There are important consequences to each of these forms of linearity. The fact that the conditional mean function is linear in the predictor variables seems to imply that linear models are only of use when the phenomenon under study is linear – yet we know that many processes in the real world do not behave linearly. However this is not as big of a problem as it may seem. Linear regression can easily accommodate some forms of nonlinearity by including nonlinear functions of the predictor variables as additional predictor variables. For example, if we represent the conditional mean function as

$$E[y|x_1, x_2] = b_0 + b_1x_1 + b_2x_1^2 + b_3x_2 + b_3x_1x_2,$$

we have added the quadratic term x_1^2 and the interaction term x_1x_2 as predictor variables, in addition to original predictor variables x_1 and x_2, which here would be called main effects. This allows us to capture conditional mean functions that are not linear functions of the predictor variables. Thus, sense (i) of linearity given above is not actually a requirement of a linear statistical model.

The second type of linearity, denoted (ii) above, has completely different implications. The fact that the conditional mean function is linear in the regression parameters (b_0, b_1, ...) is what allows us to use linear least squares to estimate these parameters from the data. In fact, the term "linear" in "linear least squares" is usually taken to refer only to sense (ii) of linearity, not to sense (i). We will provide more details about linear least squares below. At this point, it is only important to understand that the conditional mean function must be a linear function of its regression parameters if we plan to use linear least squares for estimation.

Variation

Up to this point, we have focused on the conditional mean function as the primary object of interest in a regression analysis. Almost of equal importance, however, is the conditional variance function, which we will denote $\text{Var}[y|x_1, \ldots, x_p]$. The conditional variance function quantifies the degree of scatter in the data around the conditional mean function. For example, in our blood pressure study, we may have $E[y|x_1 = 35, x_2 = 25] = 130$ – that

is, the average blood pressure for a 35 year old person with BMI equal to 25 is 130 mm/Hg. As noted above, this will not be the exact blood pressure of any individual 35 year old person with BMI equal to 25. Individual people (with that given age and BMI) will almost always have blood pressure that is either higher or lower than the conditional mean. The conditional variance function quantifies the variation of the data around the conditional mean, just like the ordinary (unconditional) variance quantifies the variance of the data around an ordinary (unconditional) mean.

The conditional variance function, like the conditional mean function, depends on its arguments $(x_1, x_2, \ldots)$. However in some populations that we study with regression analysis, the conditional variance turns out to be nearly constant in these arguments. That is, the scatter around the conditional mean has similar magnitude regardless of the values of the predictor variables. In our blood pressure example, this would mean that if the variance of blood pressures in the population at age=30, BMI=25 is equal to 5 $(\text{mm/Hg})^2$, then the variance of blood pressures at age=50, BMI=30 is also 5 $(\text{mm/Hg})^2$ (and similarly is equal to 5 $(\text{mm/Hg})^2$ for every other age and BMI value). This property, known as homoscedasticity, does not need to hold in order to conduct a regression analysis. But the most common and basic methods for conducting regression analysis work best when the conditional variance function is approximately constant in this sense.

Causality

Since we view the conditional mean as being a function with the predictor variables as inputs and the dependent variable as the output, it is tempting to think of this as reflecting a mechanism in which changes in the predictor variables can cause changes in the dependent variable to happen. In general, however, statistical analysis does not support such causal interpretations. When describing the regression model with conditional mean function, say, $E[y|x_1, x_2] = 1 + 3x_1 - 2x_2$, it is usually better to avoid saying something like "when x_1 goes up by 1 unit and x_2 is held fixed, y goes up by 3 units on average." Instead, it is usually better to say "when comparing two individuals whose x_1 values differ by 1 unit, and whose x_2 values are the same, the value of y will differ on average by 3 units." Similarly, instead of saying something like "x_1 affects y" it is better to say that "x_1 is associated with y, after controlling for x_2".

We note that there are some situations where such caution is not needed. If the data were collected as part of an experiment (i.e. if the values of x_1 and x_2 were randomly assigned to the subjects), then certain types of causal statements can be made. There is also a branch of statistics known as causal inference that aims to allow causal statements to be made from non-experimental data. While the methods from this field are very useful, they usually do not allow causal statements to be made from non-experimental data without qualification.

Estimation

Any statistical model must be fit to the available data, a process often referred to as parameter estimation. If we model the population conditional mean as $E[y|x_1, x_2] = b_0 + b_1x1 + b_2x_2$, then the estimated conditional mean function will be written $\hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2$. Here, $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$ are estimated parameters, which will rarely be exactly equal to their population counterparts (e.g. $\hat{b}_1$ will differ from b_1). We aim to recover these parameters as accurately as possible from the available data.

Linear least squares is the estimation process that we will focus on here. As noted above, this approach has successfully been used to fit linear models for well over 100 years. We will not get into the calculational details here, but will note that a modern computer can easily fit very large models with linear least squares - for example, a data set with 1 million rows (cases) and 20 predictor variables can be fit in well under three seconds.

Linear least squares works best when the conditional variance (discussed above) is constant. However the estimates of the model parameters produced by linear least squares generally remain accurate when the conditional variance is not constant. More advanced approaches to regression analysis address more completely the challenges of working with non-constant variance (heteroscedasticity). Two such approaches, marginal regression and multilevel modeling, are discussed in course 3 of this specialization.

The field of statistics focuses on estimation, but also places great importance on "quantifying uncertainty" - that is, characterizing the likely degree of discrepancy between the parameter estimates and their corresponding population values. This is a big topic that we cannot consider in detail here, but we will raise a few key points. First, any statistical estimator will exhibit some combination of bias and estimation variance. Bias is usually a bad

thing, but it is not always the overriding concern – that is, in some cases we tolerate a certain amount of bias if it brings with it lower variance or some other favorable consequence. Linear least squares is usually unbiased, which is one of its favorable attributes. However even in linear least squares, bias can sometimes result due to mis-representativeness of the data relative to the population of interest, or to systematic measurement errors in the data. Nevertheless, in basic usage bias is normally not a major concern with linear regression analysis.

Estimation variance is inevitable in any statistical analysis. It reflects the fact that we can never recover a population exactly using a finite amount of data. The estimation variance is determined predominantly by the sample size – the more data we have, the lower the estimation variance will be. Estimation variance in a regression model is also strongly influenced by three other characteristics – the level of conditional variance, the variance of the predictor variables, and the correlations among the predictor variables. We discuss each of these in turn:

- The conditional variance is the "scatter" in the data around its conditional mean. The greater the scatter, the more obscured the conditional mean will be by this "noise". Thus, the greater the conditional variance, the greater the estimation variance for the regression parameters. This type of variance is "bad variance" in the sense that greater conditional variance results in greater uncertainty about the regression parameters.

- The variance of the predictor variables refers to how dispersed the values of the different predictor variables are within the data set being used to fit the model. The primary purpose of a regression model is to establish how differences in the values of a predictor variable relate to differences in the expected values of the outcome variable. For example, we may want to know about the average difference between the blood pressures of two people who are 1 unit apart in terms of BMI (and are identical in all other measured characteristics). If our dataset has a wide range of BMI values, for example, if we have people with BMI ranging from 18 to 40, then we will have a relatively easy time quantifying how changes in BMI correspond to changes in blood pressure. Conversely, if everyone in our data set has almost the same BMI, then it will be very difficult to estimate how changes in

BMI correspond to changes in blood pressure. Note that in contrast to the conditional variance discussed in the previous point, variance in the predictor variables is "good variance," since greater variance in the predictor variables results in less uncertainty about the regression parameters.

- Correlation among the predictor variables is referred to as collinearity, and plays an important role in regression analysis. This is a more difficult topic and we will only introduce it here. Using our blood pressure study again as an example, suppose that BMI and age are highly correlated in our data set – perhaps higher BMI values occur exclusively in older people, and lower BMI values occur exclusively in younger people. In this case, if the older/higher BMI subjects have higher blood pressure, and the younger/lower BMI subjects have lower blood pressure, then the linear least squares estimator will not know whether to attribute this trend to BMI or to age. On the other hand, if BMI and age are uncorrelated, then we will have a much easier time disentangling the roles of age and BMI. Note that we are not saying here that predictor variables must be perfectly uncorrelated (i.e. that they must have zero correlation). We are only saying that greater correlation among the predictor variables tends to result in greater uncertainty in the parameter estimation. This can be overcome by collecting more data, but for a fixed amount of data, our estimates of the regression parameters will be less precise when substantial collinearity is present. We can view correlations between predictor variables as "bad correlations", because they adversely impact our ability to fit a regression model. Conversely, correlations between predictor variables and the dependent variable are "good correlations", because they allow us to fit models that do a better job of explaining the variation in the dependent variable.

Explained variation

One way to view a regression analysis is as an effort to "explain the variation" in the dependent variable, using the predictors as explanatory factors. This allows us to establish a link between linear regression analysis and the more basic idea of Pearson correlation (i.e. the familiar correlation coefficient). Suppose we have fit a linear model to data using linear least squares, and have thereby obtained parameter estimates $\hat{b}_0$, $\hat{b}_1$, ... We can use these parameter

estimates to produce fitted values. To form the fitted value for a particular observation with covariate values $x_1, x_2, \ldots$, we form the linear combination $\hat{b}_0 + \hat{b}_1 x_1 + \ldots$. This expression follows the form of the population conditional mean function, substituting the parameter estimates for the true parameter values (which are not known), and substituting the predictor variable data for one specific case for the arguments of the conditional mean function.

Once we have the fitted values in-hand, we can take the Pearson correlation coefficient between these fitted values and the observed values of the dependent variable (y). The fitted values are intended to track with the dependent variable. The closer they do so, the better the apparent explanatory performance of the model. The squared Pearson correlation coefficient between the fitted values and the observed value of the dependent variable is called the R-squared, the "proportion of explained variance," or the "coefficient of determination." The R-squared falls between 0 and 1. In general a higher R-squared is seen as reflecting a better fit of the model, but this interpretation should be qualified in two ways: first, "goodness of fit" refers to more than just the mean function – to have a model that fits well, we would like to capture the variance structure as well as the mean structure; second, higher R-squared can reflect "overfitting," in which the model fits the data in-hand better than it will fit equivalent data that we observe in the future.

Statistical inference

In course 2 we discussed statistical inference for means and proportions. The discussion there focused in particular on using standard errors, confidence intervals, and hypothesis tests as three ways to quantify the accuracy of estimated population parameters. All three of these concepts can also be used in a linear regression analysis. For each estimated regression parameter, say $\hat{b}_1$, we have a standard error s_1. Roughly speaking, the value of s_1 is the average discrepancy between $\hat{b}_1$ and its population value, which is b_1. Each regression parameter estimate will have its own standard error, reflecting the unique level of information about each parameter in the data. As noted above, several factors influence the uncertainty in a parameter estimate, including primarily: (1) sample size, (2) conditional variance in the dependent variable, (3) variance of the predictor variable, and (4) collinearity. Note that factors 1-2 impact all parameters equally, while factors 3-4 impact different parameters in a model to different extents.

Once the parameter estimates are in-hand, we can construct 95% confidence intervals for each regression parameter. For example, if we estimate that average blood pressure differs by 0.8 mm/Hg between two people who are one year different in age (and who are similar in all other measured ways), and if we have obtained a standard error of 0.3 for this estimate, then the 95% confidence interval will span roughly from 0.2 to 1.4 (+/- two standard errors from the point estimate). This can be taken to mean that although the regression parameter was estimated to be 0.8, any value between 0.2 and 1.4 would be consistent with the observed data.

Another common statistical inference task that arises in regression analysis is hypothesis testing for single parameters. For example, we might state a null hypothesis that blood pressure is unrelated to age, at a fixed BMI. This is equivalent to stating that the regression coefficient for age is equal to zero. For the point estimate and standard error given above, we would obtain a p-value for this hypothesis of around 0.008, indicating fairly strong evidence against the null hypothesis.

In addition to the statistical inference tasks described above, which are closely analogous to statistical inference tasks seen in course 2 for the setting of analyzing means and proportions, there are some additional statistical inference tasks that are more specific to regression analysis. One such task is an omnibus test for the structure of the model, which generalizes the single-parameter hypothesis tests discussed above. For example, we could conduct an omnibus test of the null hypothesis that all the population regression slopes are zero. If we cannot reject this null hypothesis, then there is no evidence that any of the predictor variables are informative about the conditional mean of the dependent variable. This type of inferential procedure is usually accomplished using "F-tests." This is a more advanced topic that we will not discuss further here.

Categorical predictor variables

In a linear regression analysis, it is common that we would like to use categorical variables as predictor variables. For example, we may want to use a person's gender as a predictor of their blood pressure. A categorical variable is not numerical, and hence cannot be directly inserted into the linear predictor $b_0 + b_1x_1 + \ldots + b_px_p$. This issue is typically addressed by coding the categorical variable into one or more indicator variables. An indicator

variable can be constructed for each level of a categorical variable. It takes on the value 1 when a case has the given level and 0 when the case does not. For example, if we want to construct indicator variables for gender, we might have a female indicator variable that is 1 for females and 0 for males, and a male indicator variable that is 1 for males and 0 for females (if there were additional gender categories, those would also be given their own indicator variables). Similarly, if we had a categorical variable that contained a person's income quartile, this would be coded as four indicator variables, with one indicator variable for each of the four levels of the income quartile variable.

When including indicator variables as predictor variables in a regression model, one indicator derived from each parent categorical variable must be dropped. The reason for this is technical, but has to do with the fact that the sum of all indicator variables derived from the same parent variable is identically equal to 1, and hence is the same as the intercept. For the model to be estimable, we cannot include the same variable twice as a predictor variable. Omitting one indicator variable avoids this issue. The level that is dropped is called the reference level. For example, with gender we may choose to omit the male indicator variable and include only the female indicator variable as a predictor in the model. In this case, "male" is the reference category.

The interpretation of a regression parameter for an indicator variable should account for which category was selected as the reference level. The regression parameter for any non-reference level is the contrast between units with that level of the parent variable, and units with the reference level of the parent variable. For example, the coefficient for the "female" variable would be interpreted as the difference between females and males. In a model for blood pressure, we might find that the coefficient for the "female" variable is -4. This would mean that holding other factors fixed, females have on average 4 mm/Hg lower blood pressure than males.

Interactions

Linear models are a form of "additive model," since the linear predictor is a sum of contributions from distinct terms. If we consider a basic linear model with mean structure $E[y|x_1, x_2] = b_0 + b_1x_1 + b_2x_2$, we see that $E[y|x_1, x_2]$ is linear in x_1 for a fixed value of x_2, and it is linear in x_2 for a fixed value of x_1. Moreover, if we fix x_2 at different values, the slope of $E[y|x_1, x_2]$ on x_1

does not change (and similarly if we swap the roles of x_1 and x_2). To put this in the context of our example, suppose that x_1 is age, and x_2 is BMI. This additivity would mean that the rate at which blood pressure changes with age is the same for people of different BMI values (as long as the BMI value itself is not changing).

While additivity is a property of basic regression models, it may not be a property of the population under study. In this case, the model would be mis-specified relative to the population. To address this, it is possible to add "interactions" to the model. Interactions can be realized in various ways, but by far the most common form for an interaction results from taking a product between two variables, and including the product as an additional predictor variable. For example, we could model the conditional mean as $E[y|x_1, x_2] = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$. In this model, the slope of $E[y|x_1, x_2]$ on x_1 for fixed x_2 is $b_1 + b_3x_2$, and the slope of $E[y|x_1, x_2]$ on x_2 for fixed x_1 is $b_2 + b_3x_1$. Here we see that the slope on x_1 depends on x_2, and the slope on x_2 depends on x_1.

In general, when including interactions in a regression model, it is good to center the predictor variables before taking the product, i.e. subtract the mean of x_1 from every value in x_1, and similarly for x_2. After doing this, the main effects gain a clear interpretation – b_1 is the slope of $E[y|x_1, x_2]$ on x_1 when x_2 is held fixed at its mean, and b_2 is the slope of $E[y|x_1, x_2]$ on x_2 when x_1 is held fixed at its mean. The coefficient b_3 continues to describe how these slopes change when the other variable is held fixed at a value that is different from its mean value.

Goodness of fit and model diagnostics

A regression model is a fairly complex object – it is a function of several variables, that depends on multiple estimated parameters. Therefore, not all the questions we may want to address with a regression model are amenable to formal inferential procedures such as F-tests. As an alternative to such formal procedures, there are a number of more informal procedures that can be very useful for assessing the structure of a fitted regression model, and for evaluating how well it fits the data. We discuss a few of these procedures here.

Above we discussed the fitted values, here we need a related quantity called the residuals, which are simply the differences obtained by subtracting each

fitted value from its corresponding observed value. For example, if the fitted blood pressure for a 25 year old with BMI 20 is equal to 115, and a specific 25 year old with BMI 20 has blood pressure 112, then the residual for this individual is -3.

A scatterplot of the residuals against the fitted values can be very informative about the structure of the fitted model. In particular, if the degree of scatter in this plot increases (or decreases) from left to right, we have discovered a mean/variance relationship. This implies that not only is the variance not constant (i.e. there is heteroscedasticity), but also that the conditional variances differ in a way that is predictable from the mean. This is an important discovery that adds depth to our understanding of the population under study. Moreover, knowing that a mean/variance relationship is present indicates that while the linear least squares estimates of the regression parameters are meaningful, their standard errors may not be correct, and we should not rely on the results obtained from standard procedures for statistical inference in linear models. Fortunately, alternative procedures that are "robust" to this heteroscedasticity are available, but this is a more advanced topic that we will not cover further here.

Scatterplots of the residuals on individual covariates can sometimes reveal nonlinear structure that the model failed to capture. If, for example, there is a strongly curved or "U-shaped" relationship between the residuals and the values of a covariate, say x_2, then it may be advisable to include x_2^2 or some other transformed version of x_2 in the model.

An added variable plot is one type of "regression graphic" that aims to reveal the structure of the regression relationship between the dependent variable y and any one of the independent variables, while controlling for the other predictor variables. It can be used to assess whether the relationship is linear or curvilinear. We will not discuss the details of how an added variable plot is constructed, but to use an added variable plot, we can inspect the plot for signs of substantial curvature. If such features are evidence, it may be advisable to transform the data or otherwise modify the model to accommodate it. There is also a more sophisticated type of regression graphic called a CERES plot which has a similar goal and interpretation as an added variable plot, but is constructed in a different way.

A partial residual plot (also called a component plus residual plot) is a graphical method that is very useful for conveying the structure of a re-

gression model to an audience who may not understand regression analysis very deeply. A partial residual plot may be used for diagnostic purposes, but more commonly is used in a didactic manner. This plot essentially constructs a synthetic data set in which the effects of all but one variable are removed. The effect of one variable of interest (called the focus variable) is not removed, and the unexplained variation is not removed either. This allows us to construct a scatterplot in which the explained and unexplained variation of one variable are visually evident (the explained variation is reflected in the trend line of this plot, and the unexplained variation is reflected in the scatter around this trend line). The main value of this plot is to overcome the challenges of visualizing a regression function, which is generally a higher-dimensional function that cannot be presented as a single scatterplot.