Regression — the basics

When we speak of *regression data*, what do we mean? The regression framework is characterized by the following:

(1) we have one particular variable that we are interested in understanding or modelling, such as sales of a particular product, or the stock price of a publicly traded firm. This variable is called the *target*, or *dependent* variable, and is usually represented by y.

(2) we have a set of p other variables that we think might be useful in predicting or modelling the target variable (say the price of the product, the competitor's price, and so on; or the profits, revenues, financial position of the firm, and so on). These are called the *predicting*, or *independent* variables, and are usually represented by x_1, x_2, etc.

Typically, a regression analysis is used for one (or more) of three purposes:

(1) prediction of the target variable (forecasting).

(2) modelling the relationship between $\mathbf{x}$ and y.

(3) testing of hypotheses.

The basis of what we will be talking about most of the semester is the *linear model*. The model can be characterized as follows. We have n sets of observations $\{x_{1i}, x_{2i}, \ldots, x_{pi}, y_i\}$, which represent a random sample from a larger population. It is assumed that these observations satisfy a linear relationship,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i,$$

where the $\boldsymbol{\beta}$ coefficients are unknown parameters, and the ε_i are random error terms. By a *linear* model, it is meant that the model is linear in the **parameters**; a quadratic model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,$$

paradoxically enough, is a linear model, since x and x^2 are just versions of x_1 and x_2. Why restrict ourselves to linear models? Well, they're simpler to understand, and they're simpler mathematically; but, most importantly, they work well for a wide range of circumstances (but definitely not **all** circumstances). It's a good idea when considering this (and any) statistical model to remember the words of a famous statistician, George Box: "All models are wrong, but some are useful." We do **not** believe that the linear model represents a *true* representation of reality; rather, we think that perhaps it provides a *useful* representation

of reality. Another useful piece of advice comes from John Tukey: "Embrace your data, not your models."

Consider now a simple regression model (that is, $p = 1$). The model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

A positive value of β_1 is consistent with a *direct* relationship between x and y; e.g., higher values of height are associated with higher values of weight, or lower values of revenue are associated with lower values of profit. A negative value of β_1 is consistent with an *inverse* relationship between x and y; e.g., higher price of a product is associated with lower demand, or a lower inflation rate is associated with a higher savings rate.

Before you actually start analyzing your data, you need to think about the underlying process you are examining, and what relationships are actually of interest. For example, say you were interested in the relationship between violent crime and wealth, and you had data relating to the number of people murdered in each country in 2015 and the 2015 gross domestic product (GDP) in US dollars for each country, for a sample of 100 countries. You might consider looking at the relationship between these two variables, and you would probably find that there is a reasonably strong one, but it would not be at all interesting, as it is largely reflecting the wrong thing. When we think of crime, we think of the chances of an individual suffering a crime; that is, we think of crime *rates*. Similarly, when we think of wealth, we are typically thinking of an individual's standard of living; that is, we think of *per capita* (per person) GDP. If we do not transform these variables to the proper scale, all we will be seeing is a *size* effect; countries with more people have more murders and produce higher totals of goods and services, but that doesn't imply that wealthier countries are less safe in any meaningful way (population is a so-called *lurking variable*, or *confounding variable*). In fact, even per capita GDP is not really correct, since it does not correct for differences across countries in cost of living and inflation, so a more appropriate measure is the GDP (at purchasing power parity) per capita. Similar uninteresting relationships occur in variables that are measured at repeated time points that are not corrected for population changes or inflation effects over time. It is the responsibility of the data analyst to get the data in a form that aligns with the actual underlying process of interest as best he or she can before even starting the analysis.

Once you've decided the right way to think about the underlying process (and thereby the right variables to examine), the first step in any analysis is to **look** at the data; in the regression context, that means looking at histograms and a scatter plot. Estimating the

unknown parameters β_0 and β_1 corresponds to putting a straight line through the point cloud in the scatter plot. In order to do this, we need a rule, or criterion, that will give a reasonable line. The standard approach is *least squares regression*, where the estimates are chosen to minimize

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

This is a standard calculus problem, and was solved for the first time either by Legendre in 1805, or by Gauss in 1794 (Legendre published first, but Gauss claimed priority). It can be shown that the least squares estimates satisfy

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

There's no need to memorize these formulas; we'll use the computer to calculate them for particular data sets. It **is** worth noting one implication of them, however. Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$; that is, the *fitted value* for the the i^{th} observation as implied by the fitted regression model. Then substituting into the formulas above gives

$$\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x}).$$

That is, the regression fit implies that the best guess for an observation whose x value is one unit above the mean of x is that its y value will be $\hat{\beta}_1$ units above the mean of y. This slope coefficient gives a direct representation of how relative position in the x space relates to relative position in the y space. The difference between the observed value y_i and the fitted value $\hat{y}_i$ is called the *residual*.

An interesting historical aside is that this fact accounts for the name of the method. Sir Francis Galton, the great British biologist, gathered data about the heights of parents and their children, and noted that the coefficient $\hat{\beta}_1$ in the model above was positive, but less than one; that is, taller parents had taller children, but they were less tall than themselves, while shorter parents had shorter children, but they were less short than themselves. Galton called this "regression to mediocrity," and the term regression eventually came to be applied to all analyses of this type. *Regression to the mean* (the modern term for this effect) is ubiquitous, and often leads to mistaken impressions about the effectiveness of interventions. For example, the institution of tougher laws might appear to have the desired effect of lowering crime rates, but if the laws were originally passed in response to

unusually high crime rates, the rates would naturally fall in later time periods because of regression to the mean, whether or not new laws were passed. Similarly, if you don't feel well and go to the doctor, chances are you'll feel better in the days afterwards no matter what the doctor does because of the natural tendency to get back to your "usual" level. This is not the same as the so-called *placebo effect*, which refers to the tendency for people to feel better after getting any treatment (even if it is completely ineffective) because of psychological effects, since regression to the mean will occur even if you don't go to the doctor at all.

The least squares regression coefficients have very specific meanings. That is,

$\hat{\beta}_1$: the estimated expected change in the target variable associated with a one unit change in the predicting variable. Note the word *estimated* — remember that $\hat{\beta}_1$ is an **estimate** of β_1, not the value itself. Note also the word *associated* — we cannot say that a change in the target variable is **caused** by a change in the predictor, only that they are associated with each other (the phenomenon of lurking variables noted earlier is one way you can see a strong association that does not in any way imply causation). This is the premise of the well-known admonition that "correlation does not imply causation."

$\hat{\beta}_0$: the estimated expected value of the target variable when the predictor equals zero. Note that this might not have any physical interpretation, since a zero value for the predictor might be meaningless, or you might have no data in your sample with predictor values near zero (so interpreting β_0 would correspond to unrealistic extrapolation, which is never a good idea). In such circumstances, there is no reason to discuss $\hat{\beta}_0$, since it does not have any practical meaning. For this reason people sometimes center a predicting variable so that the zero value is meaningful.

Note that since the regression coefficients are in the same units as the response and (for the slope) predictor variables, it is a good idea to make those units sensible. So, for example, if a variable has a range of \$1 million to \$100 million, it is much more reasonable to define it in terms of units of millions of dollars rather than dollars, since (for example) a change of \$1 is likely to be unimportant in a practical sense while a change of \$1 million could be important.

Who says that least squares regression is a good idea? Nobody, unless we make certain assumptions about our data set. We already mentioned one — the linear model should be appropriate. We need a few more assumptions in order to justify using least squares regression:

(a) the expected value of the errors is zero ($E(\varepsilon_i) = 0$ for all i). That is, it cannot be true that for certain subgroups in the population the model is consistently too low, while for others it is consistently too high. A violation of this assumption will lead to difficulties in estimating β_0, and means that your model does not include a necessary systematic component.

(b) the variance of the errors is constant ($V(\varepsilon_i) = \sigma^2$ for all i). That is, it cannot be true that the model is more accurate for some parts of the population (smaller σ) and less accurate for other parts (larger σ). This property is called *homoscedasticity*, and its violation is called *heteroscedasticity*. A violation of this assumption means that the least squares estimates are not as efficient as they could be in estimating the true parameters, and better estimates can be calculated. It also results in poorly calibrated prediction intervals.

(c) the errors are uncorrelated with each other. That is, it cannot be true that knowing that the model underpredicts y for one particular case tells you anything at all about what it does for any other case. This violation most often occurs in data that are ordered in time (time series data), where errors that are near each other in time are similar to each other (such time–related correlation is often called *autocorrelation*). Violation of this assumption can lead to very misleading assessments of the strength of the regression.

(*) the errors are normally distributed. This is needed if we want to do any confidence or prediction intervals, or hypothesis tests, which we usually do. If this assumption is violated, hypothesis tests and confidence and prediction intervals can be very misleading.

It can be shown that if these assumptions hold, least squares regression is the "right" thing to do. We will spend a lot of time this semester talking about how to check these assumptions, and how to address problems if they don't hold.

The table on the next page summarizes the assumptions, and problems associated with their violation.

Assumption	**What does it really mean?**	**When is it likely to be violated?**	**Why is it a problem?**
$E(\varepsilon_i) = 0$ for all i	It cannot be the case that some members of the population have y value that is systematically below the regression line, while others have y value systematically above it.	Well–defined subgroups in the data can cause this problem. For example, if $x \equiv$ Years on the job, and $y \equiv$ Salary, and and MBAs make more than non-MBAs, they will have $E(\varepsilon_i) > 0$, while the non-MBAs have $E(\varepsilon_i) < 0$.	Estimates of β_0 will be inappropriate. More importantly, a part of the signal is being mistakenly treated as noise.
$V(\varepsilon_i) = \sigma^2$ for all i (homoscedasticity)	It cannot be the case that the x/y relationship is stronger for some members of the population, and weaker for others (hetero-scedasticity).	Well–defined subgroups in the data can cause this problem. For example, it could be the case that the salaries of MBAs vary less around their typical values than those of non-MBAs. Another possible cause is if the data vary over a wide range. Say, e.g., that $y \equiv$ Revenues of a firm, while $x \equiv$ the Advertising budget. It is reasonable to expect that it would be possible to predict revenues more accurately for smaller firms than for larger ones.	Estimates of β_0 and β_1 will be less accurate than they could be. More importantly, assessments of predictive accuracy will be incorrect.
ε_i and ε_j are not correlated with each other for $i \neq j$.	It cannot be the case that knowing that the value of y for the i^{th} case is, e.g., below its expected value tells us anything about whether the value of y for another case is above or below its expected value.	This occurs most often for time series data. It is quite likely that if, e.g., sales of a product are higher than expected in July, they will also be higher than expected in June and August.	Measures of the strength of the relationship between x and y can be very misleading.
$\varepsilon_i \sim N(0, \sigma^2)$	The errors are normally distributed.	Can happen any time.	Confidence and prediction intervals, and hypothesis tests, can be misleading.

How can we evaluate the strength of the observed regression relationship? It can be shown that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

$$\text{Corrected total SS} = \text{Residual SS} + \text{Regression SS}$$

$$\text{Variability before regression} = \text{Variability after regression} + \text{Variability due to regression}$$

This says that the variability in the target variable can be split into two parts — the variability left over after doing the regression, and the variability accounted for by doing the regression. This immediately implies that a good regression is one with a large R^2, where

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \equiv \frac{\text{Regression SS}}{\text{Corrected total SS}}.$$

The R^2 value (also called the coefficient of determination) measures the proportion of variability in y accounted for by the regression. Values closer to 1 indicate a strong regression, while values closer to 0 indicate a weaker one. Sometimes a slightly adjusted value of R^2, which is designed to offset an upwards bias in it, is reported; the *adjusted* R^2 has the form

$$R_a^2 = R^2 - \frac{1}{n-2}(1 - R^2).$$

Is there a significant relationship between x and y? This can be tested using the F–statistic. The hypotheses being tested are

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0.$$

The test statistic is then

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2/1}{\sum(y_i - \hat{y}_i)^2/(n-2)} \equiv \frac{\text{Regression MS}}{\text{Residual MS}},$$

where MS refers to "mean square," the sum of squares divided by its degrees of freedom. The F–statistic is compared to an F–distribution on $(1, n-2)$ degrees of freedom.

Hypotheses of this type can also be tested using t–tests. To test

$$H_0 : \beta_j = \beta_j^0$$

versus

$$H_A : \beta_j \neq \beta_j^0,$$

use the statistic

$$t = \frac{\hat{\beta}_j - \beta_j^0}{s.e.(\hat{\beta}_j)},$$

substituting in the appropriate j and β_j^0, referring the statistic to a t–distribution on $n-2$ degrees of freedom. In simple regression the overall F–test is equivalent to the t–test for whether the slope equals 0 ($F = t^2$, and the p–values of the two tests will be identical). The t–distribution also provides the way to construct a confidence interval for a regression coefficient; a $100 \times (1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{\alpha/2}^{n-2} s.e.(\hat{\beta}_j),$$

where $t_{\alpha/2}^{n-2}$ is the appropriate t–based critical value.

F– and t–tests provide information about statistical significance, but they can't say anything about the practical importance of the model. Does knowing x really tell you anything of value about y? This isn't a question that can be answered completely statistically; it requires knowledge and understanding of the data. Statistics can help, though. Recall that we assume that the errors have standard deviation σ. That means that, roughly speaking, we would expect to know the value of y to within $\pm 2\sigma$ after doing the regression (since the errors off the regression line are assumed to be normally distributed). The residual mean square

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

provides an estimate of σ^2 that can be used in this formula. Its square root ($\hat{\sigma}$) is called the *standard error of the estimate* (`Minitab` calls this `s` in the output).

An even more accurate assessment of this is provided by a *prediction interval* given a particular value of x. This interval provides guidance as to how accurate $\hat{y}_0$ is as a prediction of y for some particular value x_0; its width depends on both $\hat{\sigma}$ and the position of x_0 relative to $\bar{x}$, since values further from $\bar{x}$ are harder to predict. Specifically, for a simple regression, the standard error of a predicted value based on a value x_0 of the predicting variable is

$$s.e.(\hat{y}_0^P) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}},$$

where $\hat{\sigma}$ is the square root of the residual mean square. The prediction interval is then

$$\hat{y}_0 \pm t_{\alpha/2}^{n-2} s.e.(\hat{y}_0^P).$$

The prediction interval should not be confused with a *confidence interval* for a fitted value, which will be narrower. The prediction interval is used to provide an interval estimate for a prediction of y for one member of the population with a particular value of x_0; the confidence interval is used to provide an interval estimate for the true average value of y for all members of the population with a particular value of x_0. The corresponding standard error for a fitted value from a simple regression model is

$$s.e.(\hat{y}_0^F) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}},$$

with corresponding confidence interval

$$\hat{y}_0 \pm t_{\alpha/2}^{n-2} s.e.(\hat{y}_0^F).$$

As was noted earlier, all of these tests, intervals, predictions, etc., are based on believing that the assumptions of the regression hold. We will spend a lot of time this semester talking about how to check those assumptions, and what to do if they don't hold. Remarkably enough, a few very simple plots can provide most of the evidence you need to check the assumptions.

(1) a plot of the residuals versus the fitted values. This plot should have no pattern to it; that is, no structure should be apparent. Certain kinds of structure indicate potential problems:
   (a) a point (or a few points) isolated at the top or bottom, or left or right. In addition, often the rest of the points have a noticeable "tilt" to them. These isolated points are unusual points, and can have a strong effect on the regression. They need to be examined carefully, and possibly removed from the data set.
   (b) an impression of different heights of the point cloud as you examine the plot from left to right. This indicates heteroscedasticity.
   (c) You should **never** construct a plot of the residuals versus the (observed) response values; this will **always** exhibit a pattern unless there is no relationship between the response and predictor variable(s).

(2) if your data has a time structure to it, you should plot residuals versus time. Again, there should be no apparent pattern. If you see a cyclical structure, this indicates that the errors are not uncorrelated, as they're supposed to be.

(3) a normal plot of the residuals. This plot assesses the apparent normality of the residuals. The plot should look like a straight line (roughly). Isolated points once again represent unusual observations, while a curved line indicates that the errors are probably not normally distributed, and tests and intervals might not be trustworthy.

The matrix formulation of regression

Regression modeling also can be represented using matrices and vectors. This isn't very important for simple regression, but provides a very useful shorthand for when we generalize to more than one predictor (multiple regression).

Define the following matrix and vectors as follows:

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The regression model can then be written succinctly as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

The normal equations (which determine the least squares estimates of $\boldsymbol{\beta}$) can be shown (using multivariable calculus) to be

$$(X'X)\boldsymbol{\beta} = X'\mathbf{y},$$

which implies that the least squares estimates satisfy

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

The fitted values are then

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} \equiv H\mathbf{y},$$

where $H = X(X'X)^{-1}X'$ is the so–called "hat" matrix. This matrix will turn out to be a very important one, as we'll see later on.

Now consider p predictor variables that can be used in a linear model to predict a target y. Let

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The equation $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ states that

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i,$$

a linear relationship. Thus, the same matrix formula still represents the linear regression model. This carries over to least squares estimation, as it turns out that $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$ and $\hat{\mathbf{y}} = H\mathbf{y}$, where $H = X(X'X)^{-1}X'$.