Lecture Notes: Simple Linear Regression

1 Introduction

Regression Analysis is one of the simplest ways we have in statistics to investigate the relationship between two or more variables related in a *non-deterministic* way. Examples of a deterministic relationship: $F = ma, C = 2\pi r$. If we know exactly everything on the right hand side of the equations, we know exactly everything on the left hand side. Examples of a probabilistic relationship: $X \sim Bern(p), Y \sim \mathcal{N}(n, p)$. If we know exactly the parameters on the right hand sides of the relationship, there are still probabilistic (or non-definite) values on the left hand side of the relationship.

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 competitors 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 x1, x2, 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 x and y; or

3. testing of hypotheses.

The basis of what we will be talking about most of the semester is the linear model. Virtually all other methods for studying dependence among variables are variations on the idea of linear regression. Why restrict ourselves to linear models? Well, they are simpler to understand, and they are 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."

2 The simple linear regression model

A simple deterministic relationship between two variables (here x and y) is a linear relationship

$$y = \beta_0 + \beta_1 x$$

You remember from algebra and geometry that this is the equation of line with a slope β_1 and a y-intercept β_0. What is the value of y when $x = 0$? How much does the value of y change for each unit change x. The generalization of the above model assumes that *the expected value of Y is a linear function of x, but that for fixed x, the variable Y differs from its expected value by a random amount*.

This is the *Simple Linear Regression Model*: With parameters $\beta_0, \beta_1,$ and σ^2 the dependent variable is related to x through the equation

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

The quantity ϵ in the model equation is a random variable (and hence Y is as well). We usually call ϵ the random error term. Without it, any observed pair (x, y) would fall exactly on the line $y = \beta_0 + \beta_1 x$. The inclusion of the random error term *scatters* the points $(x_1, y_1), ..., (x_n, y_n)$ about the true regression line. This ϵ is the *deviance* of the data from the linear model.

We are commonly interested in the behavior of β_1. 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.

Conventionally, we call x the *independent variable* and y the *dependent variable*. The dependent variable is also called *response variable*. The independent variables are also called *explanatory variables.* Depending on the objective of the regression, these variables may be also called *predictors* when the objective of the regression is prediction, *controlling factors* when the regression is used to control for biases in the experiment, *covariates* which can include both predictors and controlling factors, etc.

In the simplest scenario we have one response variable (Y) and one explanatory variable (X). For example, we might predict a son's height, based on his father's height (Figure 1). Or we might predict a cat's heart weight, based on its total body weight (Figure 2).

Implications of the regression model (equation)

Suppose (X, Y) have a joint distribution $f(x, y)$. You observe $X = x$. Then the expected or mean value of Y when we know x is:

$$\mu_{(Y|x)} = (\beta_0 + \beta_1 x + \epsilon) =$$

$$= \beta_0 + \beta_1 x + E(\epsilon) =$$

$$= \beta_0 + \beta_1 x$$

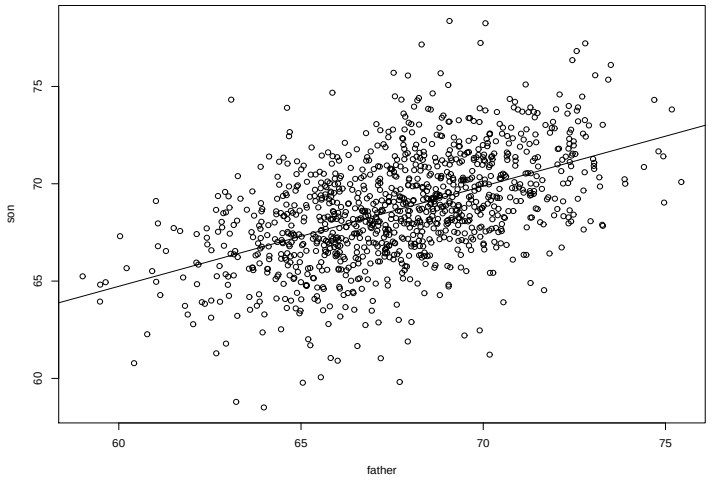

Figure 1: Galton data. Predict Son's height from Father's height.

This states that the *mean value* of Y, rather than Y itself, is a linear function of x. The true regression line $y = \beta_0 + \beta_1 x$ is called the *line of mean values*; it's height above any particular x value is the expected value of Y for that given x. The slope β_1 is interpreted as the *expected* change in Y associated with a unit increase in the value of x.

The variance of Y when we know x is:

$$\sigma^2{}_{(Y|x)} = Var(\beta_0 + \beta_1 x + \epsilon) =$$

$$= Var(\beta_0 + \beta_1 x) + V(\epsilon) =$$

$$= 0 + \sigma^2 = \sigma^2.$$

This states that the amount of variability in the distribution of Y values is the same at each different value of x (under this model). This is called *homogeneity of variance*.

Together, these equations (with the model) state that, given observed fixed x, Y is the sum of a constant $\beta_0 + \beta_1 x$ and a normally distributed random variable ϵ, so Y itself has a normal distribution.

3 Simple Linear Regression: Estimation

The parameters $\beta_0, \beta_1, \sigma^2$ are unknown that are estimated from the sample data $(x_1, y_1), ...(x_n, y_n)$, the n pairs of observations. 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 or to generate estimates of the population parameters – using the hat notation we'll call them $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2$. In order to do this, we need a rule, or criterion, that will give a reasonable line.

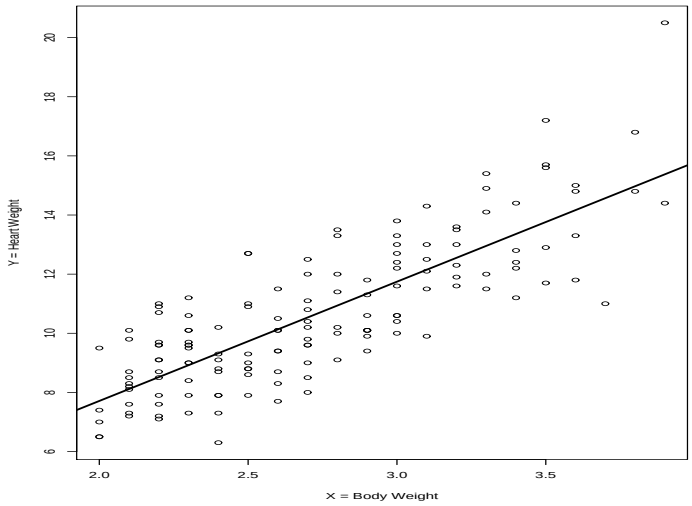

Figure 2: Cat example

Intuitively, a reasonable estimate of the regression line will "fit" the observed data. In fact, any line is completely determined by the values chosen for β_0 and β_1. So then our job is to choose the best values $\widehat{\beta}_0, \widehat{\beta}_1$ – i.e. choose the best line, i.e. choose the line which "fits" the data best. The common approach for obtaining estimates for β_0 and β_1 is to minimize the sum of squared deviations, the sum of squared errors (SSE) from the regression line.

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \\ &= \sum_{i=1}^{n} [y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2 \\ &= \sum_{i=1}^{n} (\widehat{\epsilon}_i)^2 \end{aligned}$$

For a fixed set of observations (i.e. observed values) $(x_1, y_1), ..., (x_n, y_n)$. We need to find the minimum of this SSE function, that is, we *minimize the sum of squared error* or we *minimize the sum of the squared residuals*. The SSE is a function in β_0, β_1 for a fixed set of observations $(x_1, y_1), ..., (x_n, y_n)$. We can find the extreme values of a function by taking a derivative and setting it equal to zero. 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).

Specifically, we will find the minimum of the SSE function by taking 1st order derivatives and setting the resulting equations equal to zero. Moreover, assuming normality of the data, you will recognize this procedure as another version of the *method of maximum likelihood* estimation. So pause and think to yourself that we are estimating parameters again, using a *likelihood.* Here, the likelihood is generated by the distribution on the residuals - the error in the model.

Deriving the Estimators

The function:

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \widehat{y})^2$$

has a partial derivative with respect to β_0

$$\frac{\partial SSE(\beta_0, \beta_1)}{\partial \beta_0} =$$

$$= \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)(-1) = 0$$

The partial derivative, with respect to β_1

$$\frac{\partial SSE(\beta_0, \beta_1)}{\partial \beta_1} =$$

$$= \sum_{i=1}^{n} 2(y_i - (\beta_0 + \beta_1 x_i)(-x_i) = 0$$

Canceling and rearranging both equations gives us the *normal equations*.

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i$$

$$\beta_0 (\sum x_i) + \beta_1 (\sum x_i^2) = \sum x_i y_i$$

The solutions to these equations

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

$$= \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n(\sum {x_i}^2) - (\sum x_i)^2}$$

and

$$\widehat{\beta}_0 = \frac{\sum y_i - \widehat{\beta}_1 \sum x_i}{n}$$

$$= \overline{y} - \widehat{\beta} \overline{x}.$$

We call the estimates of the y_i's y-hat's or $\widehat{y}_i$'s or the *fitted values*

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

and the *residuals* $\widehat{\epsilon}_i$'s

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$

$$= y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i).$$

We estimate the variance of the error terms with the residuals

$$\widehat{\sigma}^2 = \frac{SSE}{n-2}$$

$$= \frac{\sum (y_i - \widehat{y}_i)^2}{n-2}$$

$$= \sum (y_i)^2 - \widehat{\beta}_0 \sum y_i - \widehat{\beta}_1 \sum x_i y_i$$

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 $\widehat{\beta}_1$ 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 do not feel well and go to the doctor, chances are youll 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 do not go to the doctor at all.

Interpretation

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

- $\widehat{\beta}_1$: the estimated expected change in the target variable associated with a one unit change in the predicting variable. Note the word estimated remember that $\widehat{\beta}_1$ 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.

- $\widehat{\beta}_0$: 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 $\widehat{\beta}_0$, since it does not have any practical meaning.

Assumptions

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:

- the expected value of the errors is zero ($E(\epsilon_i) = 0$ 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.

- the variance of the errors is constant ($V(\epsilon_i) = \sigma^2$ 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.

- 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 timerelated 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 do not hold.

Analysis of Variance

The *Regression Sum of Squares* (or Sum of Squares due to Treatment) is

$$SSTR = SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

and the Total Sum of Squares is

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

.

$$= \sum (y_i)^2 - (\sum y_i)^2/n.$$

A straightforward interpretation of the variance decomposition is:
Variability before regression = Variability after regression+ 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.

Lastly, we define the *coefficient of determination* be

$$r^2 = 1 - \frac{SSE}{SST}$$

$$= \frac{SST - SSE}{SST} = \frac{SSR}{SST}.$$

It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between y and x). 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

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

These statistics are often presented in what is called an Analysis of Variance (ANOVA) table.

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	k-1(=1)	SSR	MSR=SSR	SSR/SSE/(n-2)
Error	n-k(=n-2)	SSE	SSE/(n-2)	
Total	n-1	SST		$r^2 = \frac{SSR}{SST}$

This method of estimating the parameters in the linear model is commonly known as *ordinary least squares regression*. There are many forms of regression[1] and many extensions of the linear model.[2]

3.1 Properties of Parameter Estimators

The estimators are random variables and have the following properties (**conditional on** $X_1, \ldots, X_n$):

$$\mathbb{E}(\widehat{\beta}_0) = \beta_0, \quad \mathbb{E}(\widehat{\beta}_1) = \beta_1,$$

$$\mathbb{V}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right), \quad \mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}, \quad \mathsf{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\sigma^2 \frac{\bar{x}^2}{S_{xx}},$$

where $S_{xx} = \sum_{i=1}^n (X_i - \overline{X})^2$.

Let's derive some of these facts. Let

$$d_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

[1] Remember to 'regress', in statistics, is to find an expected value for a Y given an observed value of $X = x$.

[2] In general $Y_{X=x} = f(x, \beta)$ and f – called the 'link' function – can be drawn from many families of functions.

So that $\widehat{\beta}_1 = \sum_i d_i(Y_i - \bar{Y})$. We use this fact repeatedly in the derivation. Then

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_1) &= \mathbb{E}\sum_{i=1}^{n} d_i(Y_i - \bar{Y}) \\
&= \sum_{i=1}^{n} d_i\mathbb{E}(Y_i) - \mathbb{E}(\bar{Y})\sum_{i=1}^{n} d_i \\
&= \sum_{i=1}^{n} d_i\mathbb{E}(\beta_0 + \beta_1 X_i) - \mathbb{E}(\beta_0 + \beta_1\bar{X})\sum_{i=1}^{n} d_i \\
&= \sum_{i=1}^{n} d_i[\beta_1(X_i - \bar{X})] \\
&= \beta_1\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= \beta_1
\end{aligned}$$

$$\begin{aligned}
\mathbb{V}(\widehat{\beta}_1) &= \mathbb{V}\sum_{i=1}^{n} d_i(Y_i - \bar{Y}) \\
&= \sum_{i=1}^{n} d_i^2\mathbb{V}(Y_i) \\
&= \frac{\sigma^2}{s_x^2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(\widehat{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \bar{X}\mathbb{E}[\widehat{\beta}_1] \\
&= \beta_0 + \beta_1\frac{1}{n}\sum_{i=1}^{n}(X_i) - \bar{X}\mathbb{E}[\widehat{\beta}_1] \\
&= \beta_0
\end{aligned}$$

Also, $\mathbb{E}(\widehat{\sigma}^2) = \sigma^2$. The **standard error** of $\widehat{\beta}_1$ is the standard deviation of $\widehat{\beta}_1$ evaluated using $\widehat{\sigma}^2$ for σ^2:

$$\widehat{\mathsf{se}}(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}.$$

Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear combinations of $Y_1, \ldots, Y_n$, if data are normally distributed then the estimators are also normally distributed.

Approximate Normality

$$\widehat{\beta}_0 \sim N\Big(\beta_0, \widehat{\mathsf{se}}^2(\widehat{\beta}_0)\Big), \quad \widehat{\beta}_1 \sim N\Big(\beta_1, \widehat{\mathsf{se}}^2(\widehat{\beta}_1)\Big) \tag{1}$$

If $\epsilon_i \sim N(0, \sigma^2)$ then:

1. Equation (1) is exact.
2. The least squares estimators are the maximum likelihood estimators.
3. The variance estimator satisfies:
$$\widehat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-2}}{n-2}$$
And $\mathbb{E}[\widehat{\sigma}^2] = \frac{\sigma^2(n-2)}{n-2} = \sigma^2$.

Note: To verify these results again assume calculations are performed conditional on $X_1, \ldots, X_n$. Then

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

The likelihood is

$$f(y_1, \ldots, y_n; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i; \beta_0, \beta_1, \sigma^2).$$

If we write out the likelihood of the normal model, (2) follows directly. For (3), note that $\sum_i \epsilon_i^2/\sigma^2 \sim \chi^2_n$. We can show that $\sum_i \widehat{\epsilon}_i^2/\sigma^2 \sim \chi^2_{n-2}$. Hence, $\mathbb{E}[\text{RSS}/\sigma^2] = n-2$, and the result follows.

3.2 Inference

The estimates of the coefficients and their standard errors are generally available on regression output. Tests and confidence intervals for the coefficients will use this statistic

$$T = \frac{\widehat{\beta}_i - \beta_i}{s.e.(\widehat{\beta}_i)}$$

$$s.e.(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{S_{xx}}}$$

and

$$s.e.(\widehat{\beta}_0) = \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

which has a t distribution with $n-2$ degrees of freedom, since we are using the plug in variance estimator — yielding a t-distribution.

3.3 Prediction

The 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 again! 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 provides an estimate of σ^2 that can be used in this formula. Its square root is called the standard error of the estimate.

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 y_0 is as a prediction of y for some particular value x_0; its width depends on both σ^2 and the position of x_0 relative to $\bar{x}$, since values further from $\bar{x}$ are harder to predict.

Recall that $y = \beta_0 + \beta_1 x$ yields the estimated conditional expected value of Y given $X = x_0$,

$$\widehat{\mathbb{E}(Y|X = x_0)} = \mathbb{E}(Y_{x_0}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_0.$$

This estimator of the conditional expected value is *unbiased*

$$\begin{aligned}\mathbb{E}(\widehat{E(Y|X = x_0)}) = \mathbb{E}(\widehat{\mathbb{E}(Y_{x_0})}) &= \mathbb{E}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \\ &= \beta_0 + \beta_1 x_0\end{aligned}$$

with variance

$$\begin{aligned}\mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) &= \mathbb{V}(\overline{y} + \widehat{\beta}_1 \cdot (x_0 - \overline{x})) \\ &= \sigma^2[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}]\end{aligned}$$

So a *confidence interval for the conditional expectation* is

$$\mathbb{E}(Y|X = x_0) \in \widehat{y}_{x_o} \pm t_{\alpha/2, n-2} \cdot \widehat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$

A *prediction interval* is an interval that has a probability, say $(1 - \alpha)$ of a future value — say Y^* — for an observed $X = x_0$. We estimate this future value $Y*$ with the regression equation $\widehat{Y}$.

The expected value of the difference between the prediction of the future value and the future value is zero:

$$\begin{aligned}\mathbb{E}(\widehat{Y}^* - Y^*) = \mathbb{E}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0 + \epsilon^*)) = \\ = 0\end{aligned}$$

The variance of the difference is:

$$\begin{aligned}\mathbb{V}(\widehat{Y}^* - Y^*) &= \mathbb{V}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - \epsilon^*) \\ &= Var(\overline{Y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_0 - \epsilon^*) \\ &= \frac{\sigma^2}{n} + (x_0 - \overline{x})^2 \frac{\sigma^2}{S_{xx}} + \sigma^2.\end{aligned}$$

The statistic is

$$T = \frac{\widehat{Y}^* - Y^*}{\sqrt{\frac{\widehat{\sigma}^2}{n} + (x_0 - \overline{x})^2 \frac{\widehat{\sigma}^2}{S_{xx}} + \widehat{\sigma}^2}}$$

and is distributed t_{n-2}. This yields a $(1-\alpha)\%$ *prediction interval for a single response*

$$Y^* \in \widehat{y^*} \pm t_{\alpha/2} \cdot \sqrt{\frac{\widehat{\sigma}^2}{n} + (x_0 - \overline{x})^2 \frac{\widehat{\sigma}^2}{S_{xx}} + \widehat{\sigma}^2}.$$

Remark: 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.

4 Model Assessment

4.1 R-squared

The R^2 statistic, for a linear regression between two variables, is just the square of the correlation.

$$R^2 = r^2 = \frac{cov(x,y)^2}{S_{xx}S_{yy}} \tag{2}$$

with a little algebra we can show that

$$R^2 = r^2 = 1 - \frac{\sum(y_i - \widehat{y}_i)^2}{\sum(y_i - \overline{y})^2} \tag{3}$$

This second equation is true in general. We also call R^2 the *coefficient of determination*. Looking at the second equation we see that R^2 as 1 minus the ratio of deviation between the observations and the model to the deviation between the observation and their sample mean.

We see R^2 is high when the SSE, or the deviation between the observations and the model, is low when compared to the deviation between the observations and their sample mean.[3]

4.2 Correlation and Regression

The notions of linear correlation and regression are closely intertwined. Regression is a linear relationship based on correlation – the measure of the strength of the linear relationship.

[3]What's going on? We are partitioning the error between the observations and their mean: $SST = \sum(y_i - \overline{y})^2 = \sum(y_i - \widehat{y}_i)^2 + \sum(\widehat{y}_i - \overline{y})^2 = SSE + SSreg$. The last summand is called the Sum of Squares for Regression — the error induced by fitting the regression line. We've seen the SSReg as a line in our regression output already.

What we do when we fit a regression (i.e. estimate the parameters for a regression line; the slope and the intercept) is characterize the linear relationship by drawing a line — this is called 'fitting' a regression, or fitting a regression line. A regression line is the 'best fit' line in the sense that it is the line which minimizes the difference between the observed values of the dependent variable and the predictions given the independent variable.

We should never stop at just fitting the line — very few relationships are perfect lines. We should always examine the 'fit' of the line. We do this by considering:

- Is the relationship is actually linear?
- Are the assumptions about the error in the regression are satisfied?
- Outliers and influential observations?

These are the same exercises we should perform on any model. Does our model capture the functional relationship reasonably well? Are the assumptions about the error in the model (distributional assumptions) valid? Where does the model fail - and why?

4.3 Diagnostics

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 do not hold. Remarkably enough, a few very simple plots can provide most of the evidence you need to check the assumptions.

The theme of diagnostics in the linear regression model is departures from the assumptions, which are chiefly:
1. Linear relationship between response and predictor;
2. Normality of the error;
3. Constant error variance (*homoscedasticity*); and
4. Independence of the error.

Checking Observations: Predictors and Response. Any data point that is far from the majority of the data (in both x and y) is called an *outlier*. An *outlier* is an observation that is unusually small or large.

Remember that our estimates for the regression parameters came from our attempt to fit the mean line through the y's at each value of x.

Data points that are far from the mean of the x's are called *leverage points*. A data point that is far from the the mean of both the y's and the x's are often *influential* and can change the value of the estimated parameters significantly.

The upshot: Often a regression line will differ greatly depending upon what *subsets* of the data are included. Sometimes there are good reasons for excluding subsets (there were errors in the data entry; there were errors in the experiment). Sometimes - the outlier belongs in the data. Outliers should always be examined.

4.4 Residuals

Remember our assumption (or *hypothesis*) about the error variable $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We suppose that the error has a zero mean and some known, constant variance. Our sample estimate for σ^2 is the square root of the sum of squared error (SSE) divided by $n - 2$:

$$s^2{}_e = \frac{SSE}{n-2} = \frac{\sum {e_i}^2}{n-2} = \frac{\sum (y_i - \widehat{y}_i)^2}{n-2} \tag{4}$$

Notice that we compute $\widehat{y}_i$ by substituting x_i into the formula of the regression line. The observed value of ϵ is e – we call it the residuals. The square root of our estimate for the error in the model, s_e is called the *standard error of the estimate*.

The smallest value that s_e can assume is 0, which occurs when $SSE = 0$, that is, when all the points fall on the regression line. Thus when s_e is small, the fit is good, and the linear model is likely to be an effective tool. If is s_e is large, the linear model is unlikely and we should discard or modify it.

We should always, at least graphically, check the assumptions for the residuals:

- The residuals are approximately normally distributed
- The mean of the residuals is approximately zero
- The standard deviation of the residuals is relatively constant for regardless of the value of x.
- The values of the residuals appear to be independent of y – there are no patterns

An approach to assessment of model adequacy is to compute the fitted values, $\widehat{y}_i$'s, and the residuals $e_i = y_i - \widehat{y}_i$ and then plot various functions of these computed quantities. Remember that each residual has expected value zero. If we look at the *standardized residuals*

$$e*_i = \frac{y_i - \widehat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum (x_j - \overline{x})^2}}}$$

Standardized residuals bigger than one are large; bigger than two extremely large. Most statistics packages will calculate these automatically.[4].

4.5 Diagnostic Plots

The basic plots for an assessment of model validity and usefulness are

1. x vs. $e*$ or e is called a *residual plot* against the predictor. Here, evidence **against** the regression is any non random pattern indicating a change in variance at levels of x.

[4]We are skipping the details, but the variance of the residuals (in this case the square of a standard error –why?) is $V(e_i) = \sigma^2[1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{\sum (x_j - \overline{x})^2}]$

2. $\widehat{y}$ vs. $e*$ or e is called *residual plot* against the response. Here, again, evidence **against** the regression is any non random pattern indicating a change in variance at levels of y. Both of the above are instances of *heteroscedasticity*.

3. y vs. $\widehat{y}$ is a plot of the fitted vs. the observed values. Evidence against the regression is a pattern other than a 45 degree line.

4. A plot of the quantiles of the standard normal distribution vs. $e*$, the standardized residuals. Evidence against the regression is a pattern other than a 45 degree. This plot shows, implicitly, how the residuals vary with x and how fitted values compare with observed values.

4.6 F Test for 'fit' of the model

Recall the ANOVA table is

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	MSR=SSR	SSR/SSE/(n-2)
Error	n-2	SSE	SSE/(n-2)	
Total	n-1	SST		$r^2 = \frac{SSR}{SST}$

The value in the last column

$$f = SSR/SSE/(n-2)$$

has an F distribution with 1 numerator and $n-2$ denominator degrees of freedom.

This statistic tests, in simple linear regression, the null hypothesis $H_0 : \beta_1 = 0$ — i.e., there is no useful relation between the predictor and the response.

$$H_0 : \beta_i = 0, \forall i \geq 1$$

vs.

$$H_1 : \beta_i \neq 0, \exists i \geq 1$$

An alternate way of writing the test statistic

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

Essentially the F statistic is the ratio the mean square error due to the regression (MSR) and the mean squared error (MSE).

The error sum of squares $SSE = \sum(y_i - \widehat{y})^2$ can be partitioned as

$$SSE = SSPE + SSLF$$

$$\sum_j \sum_i (y_{ij} - \widehat{y}_{ij})^2 = \sum_j \sum_i (y_{ij} - \overline{y}_j)^2 + \sum_j \sum_i (\overline{y}_j - \widehat{y}_{ij})^2$$

where we partition the predictor X over its observed levels j into c *repeated measurements*. The SSE is partitioned into a *pure error* components — $SSPE$ and a lack of fit component — $SSLF$. We can rewrite the ANOVA table with this additional partitioning...

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	MSR=SSR	SSR/SSE/(n-2)
Error	n-2	SSE	SSE/(n-2)	
Lack of Fit	c-2	SSLF	MSLF=SSLF/(c-2)	f*=MSLF/MSPE
Pure Error	n-c	SSPE	MSPE=SSPE/(n-c)	
Total	n-1	SST		$r^2 = \frac{SSR}{SST}$

The statistics f^* has F distribution with $c-2$ numerator and $n-c$ degrees of freedom. For both f statistics we reject the null hypothesis for large values of the test statistic.

For f the null hypothesis is $H_0 : Y_i = \beta_0 + \epsilon_i$; a rejection of the null indicates that there is a conditional relationship and that the predictor is useful

For f^* the null hypothesis is $H_0 : Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$; a rejection of the null indicates that the regression function/conditional expectation is not linear.