Linear Regression with One Predictor
Chapter 3

Regression is about modeling relationships between variables. The simplest kind of relationship is the linear relationship. And the simplest kind of linear relationship involves only two variables. One variable is called the **response variable** or **dependent variable** and is labeled *Y*. The other variable is called the **predictor variable** or **independent variable** and is labeled *X*. The relationship is thought to flow from *X* to *Y*, rather than from *Y* to *X*, and it is important to correctly identify the direction of the influence in the issue at hand. Understanding linear regression with one predictor is key to understanding more complex relationships.

The linear regression model with one predictor variable.

Suppose that $x_1, x_2, ..., x_n$ are given constants. Let $Y_1, Y_2, ..., Y_n$ be random variables such that for all *i* = 1, …, *n*

(1) $E(Y_i) = \alpha + \beta x_i$

(2) $Var(Y_i) = \sigma^2$

(3) $Y_1, Y_2, ..., Y_n$ are independent

(4) Y_i is normally distributed.

If all of these assumptions hold, then the linear regression model is said to be valid.

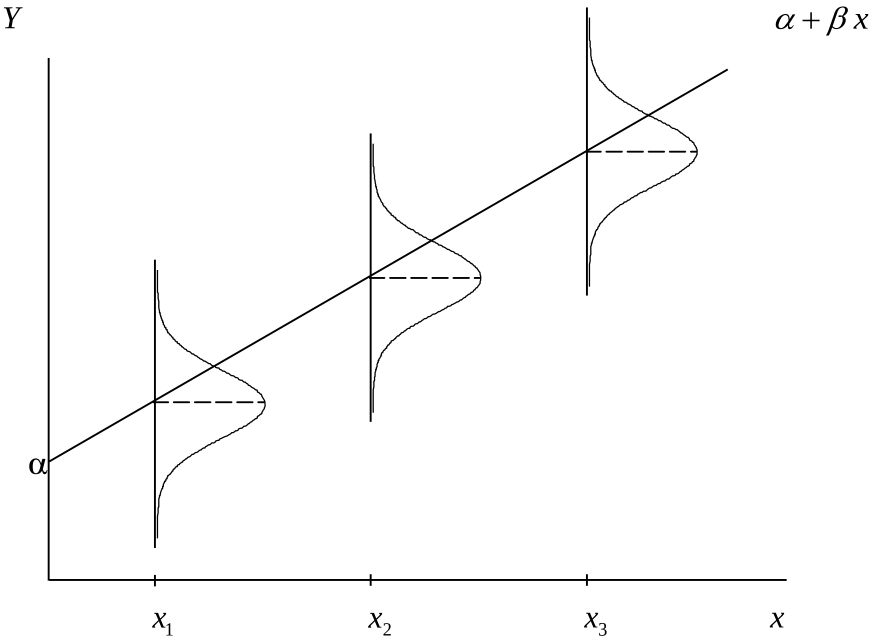
Discussion.

The four assumptions are model **specifications**:
Assumption (1) is the **linearity specification**. (L)
Assumption (2) is the **homoscedasticity specification**. (H)
Assumption (3) is the **independence specification**. (I)
Assumption (4) is the **normal specification**. (N)

The linear regression model applies exactly in very few real data analysis situations. There is usually something wrong with one or more of the specifications. However, the model often applies approximately. Moreover, specification problems can often be fixed by relatively simple means. Much of chapters 6, 8, and 9 is about detecting and fixing specification problems.

Do not be concerned about the apparent disparity between my short list of model assumptions and the long list of assumptions given by your author in chapter 3. Ramanathan collects his regression assumptions in Table 3.2. The disparity is only apparent. His long list is equivalent to my short list above.

In short, the model specifications say that at each *x*, there is a normal distribution of potential (*Y*) outcomes, and the mean of the normal distribution varies linearly with *x* but the variance does not vary with *x*, and the outcomes are all drawn independently. The following graph illustrates matters for three illustrative *x*-values:



Since the mean of *Y* depends upon the *x* at which *Y* is observed, then $Y_1, Y_2, ..., Y_n$ are not identically distributed, although they are independent. The linearity specification means *linear in the parameters*. So $E(Y_i) = \alpha + \beta x_i^2$ would still be linear (the mean of *Y* is a linear function of x^2), whereas $E(Y_i) = \alpha + \beta^2 x_i$ would not be linear.

"Homoscedasticity" is a sesquipedalian term for *constant variance*. That is, the variability of the potential outcomes for *Y* is the same, regardless of the *x* at which the *Y*'s are observed.

The model specification can be rephrased usefully in terms of model errors: Suppose that $x_1, x_2, ..., x_n$ are given constants. Let $Y_1, Y_2, ..., Y_n$ be random variables such that for all *i* = 1, …, *n*

(1) $Y_i = \alpha + \beta x_i + \varepsilon_i$ and $E(\varepsilon_i) = 0$

(2) $Var(\varepsilon_i) = \sigma^2$

(3) $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are independent

(4) ε_i is normally distributed.

If all of these assumptions hold, then the linear regression model is said to be valid.

This regression model specification is equivalent to the earlier model specifications (1) – (4). (Why?) So the regression model specification is that the random variables *Y* are a linear function of *x* plus zero-mean errors that are a normal Random Sample. This version of the specification makes it convenient to test for violations of regression model specifications. Under the null hypothesis that the four specifications hold, the errors must be a normal Random

Sample. One can therefore calculate the regression residuals as estimates of the errors and use the residuals to test the specifications.[1]

Note that the model posits that $x_1, x_2, ..., x_n$ are constants, rather than random variables. In experiments, the researcher usually can choose the *x*-values at which the *Y*'s will be observed. But in most observational settings, the researcher can not choose the values of $x_1, x_2, ..., x_n$ - they are produced by some kind of random process. Suppose that $x_1, x_2, ..., x_n$ are actually realizations of a random process. What then? Most researchers go ahead and use ordinary regression anyway. But most of those who proceed are not aware of the issues involved in proceeding and are even unaware that there are any issues to consider! There are two legitimate ways to proceed. One way is to proceed as usual under the above linear regression model (1) – (4) but recognize that your analysis is *conditional* on the *x*-values that you observe. That is, your analysis depends upon the *x*-values that you observe, and the applicability of your model to other *x*'s may be limited. The other way is to generalize the regression model to incorporate the variability of the *x*'s explicitly. Then your analysis will extend and apply readily to other *x*'s that are not observed. The latter requires changing the regression model and changing some of the inferential procedures to treat the *x*'s as random.

> Ex: You collect rent and area data on a random sample of 60 Austin apartments. You regress rent on area: RENT = a + b*AREA. OLS regression treats your 60 areas as given constants, just as though you had announced in advance of sampling that you would target apartments with these 60 specific areas and ascertain their rents. But in fact, based on the way you sampled the apartments, your areas are just as random as your rents. You can resolve this violation of the regression model in one of two ways: (1) treat your areas as constants by interpreting the inference as being about the conditional distribution of rent given area = *x*. Then you can do OLS. But the potential practical drawback of (1) is that your inferences about rent depend upon and are restricted by the specific 60 areas that you observed. So you must be careful about making generalizations about the relationship between rent and area globally. So there is reduced scope for global statements about the relationship of rent and area. The second way is (2) treat your areas as random and take the variability of *X* into account. Then you are justified in generalizing about the relationship between rent and area among all Austin apartments. You probably want to do (2). But, along with most researchers, you will probably accept the limitations of OLS and do (1). The mistake is to do (1) and believe you are doing (2).[2]

I will not develop the second approach in any detail. But I will digress for a moment to state the revised regression model for the second case, point out its relationship to the first case, and make a few comments.

The linear regression model with random *X*. Suppose that $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ are a Random Sample of bivariate normal pairs of random variables. That is, each pair (X_i, Y_i) has a bivariate normal distribution (so Y_i depends on X_i if their correlation coefficient is not zero), and each pair is drawn independently of every other pair. Then the (generalized) linear regression model applies. That is the complete specification. It is all you need to assume and/or verify for the random *X* regression model.

[1] More on testing specifications in Chapters 6, 8, and 9.

[2] If you really want to do (2), you must model the joint distribution of rent and area. Unfortunately, it is not bivariate normal. But the logarithms of rent and area are approximately bivariate normal. If you are satisfied with using and interpreting a regression of the form log *Y* = a + b log X + e, then this is an option,

It can be shown that the assumption that $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ are a Random Sample of bivariate normal pairs of random variables *implies* the four specifications (**L**, **H**, **I**, **N**) of the standard linear regression model. [3] That is, under this bivariate normal specification, the conditional distribution of *Y* given *X* = *x* is normal with mean $E(Y \mid X = x) = \alpha + \beta x$ and constant variance σ^2 and the conditional random variables $Y_1 \mid X_1 = x_1, Y_2 \mid X_2 = x_2, ..., Y_n \mid X_n = x_n$ are independent for every realization $x_1, x_2, ..., x_n$. This is the complete linear regression specification (1) – (4) given above. But the regression analysis under the bivariate normal specification is carried out by replacing $x_1, x_2, ..., x_n$ by $X_1, X_2, ..., X_n$. To evaluate the means, variances, and other properties of OLS estimators, the *x*'s may be treated as constants. But in the bivariate normal specification, they are treated as random variables. That makes the properties of the estimators more difficult to obtain.

In the more general model of random *X*'s, you recognize that your realized $x_1, x_2, ..., x_n$ could have been different. You may want to know what your data say about the whole population. Your parameter estimates would vary if you had gotten a different set of $x_1, x_2, ..., x_n$ than you actually got. Their standard errors would vary. Corresponding tests of significance would vary, some more significant, some less. If you want your results to apply for all *x*'s, then you should use standard errors and tests of significance that take *x*-variability into account.[4] It turns out that the OLS estimators of α and β in the standard, conditional case are unbiased estimators of α and β in the more general case. However, the standard errors of the OLS estimators in the standard, conditional case are biased but consistent estimators of the general case standard errors. This means that adjustments should be made in the OLS standard errors and tests of significance unless the sample size is large. If the sample size is large in a certain technical sense, then there is little difference between the two cases.

Here is a further generalization: I have stated the linear regression model for random *X* when (*X*,*Y*) have a joint bivariate *normal* distribution. This implies the conditional regression model. Quite often, however, the distribution of *X* is not normal, although the conditional distribution of *Y* given *X* may be. If *X* is not normal, the joint distribution of (*X*,*Y*) cannot be bivariate normal. But the conditional linear regression model can still hold. (Can you give an example?) That is, assumptions (1) – (4) can still hold, given a realization $x_1, x_2, ..., x_n$ of $X_1, X_2, ..., X_n$. Then the question remains, Do you want to limit your inference to the $x_1, x_2, ..., x_n$ that you observe, or do you want it to apply generally? If the former, then proceeding with the usual conditional regression model is OK. If the latter, then you should assess the consequences for inference with random *X*'s. This may be hard to do if *X* is not normal. I will have a few further things to say about the general case, but I will not go into it in detail.

Least-squares estimation of the parameters.

Suppose that $y_1, y_2, ..., y_n$ are a realization of the random variables $Y_1, Y_2, ..., Y_n$, for which the linear regression model holds at $x_1, x_2, ..., x_n$. The regression model says that y_i "should be"

[3] But please note that the standard OLS model does *not* imply the bivariate normal regression model. For example, it is possible to have a Random Sample of *non*-normal pairs $(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)$ that satisfy **L**, **H**, **I**, and **N**. (Can you give an example?).

[4] The regression output of standard computer software like SAS is for the case of conditional *x*'s, not random *X*'s.

$\alpha + \beta x_i$. Thus, the principle of least squares says to estimate α and β by the values of α and β that minimize the **Error Sum of Squares** (**ESS**) = $\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2$. Expand ESS to obtain

$$\text{ESS} = \sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 = n\alpha^2 + 2\alpha\beta\sum x_i + \beta^2\sum x_i^{\,2} - 2\alpha\sum y_i - 2\beta\sum x_i y_i + \sum y_i^{\,2} .$$

Remember that $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are realizations, so they are known values and treated as constants. It is the parameters α and β that are unknown and that we want to choose to minimize the ESS. When graphed as a function of α and β, the ESS is a paraboloid (a three-dimensional generalization of a parabola). It looks like a mountain, either right-side up or upside down. In this case, it is an upside-down mountain and so achieves a minimum at the inverted peak. Finding the minimizing values for α and β is an exercise in multivariable calculus. One takes the partial derivatives of ESS with respect to α and β, sets the resulting expressions equal to zero, and solves:

$$\frac{\partial}{\partial\alpha}(ESS) = 2n\alpha + 2\beta\sum x_i - 2\sum y_i = 0$$
$$\frac{\partial}{\partial\beta}(ESS) = 2\alpha\sum x_i + 2\beta\sum x_i^{\,2} - 2\sum x_i y_i = 0$$

So the minimizing values for α and β are the solutions to the simultaneous linear equations

$$n\alpha + \beta\sum x_i - \sum y_i = 0$$
$$\alpha\sum x_i + \beta\sum x_i^{\,2} - \sum x_i y_i = 0$$

These equations are called the **Normal Equations**.[5]

A variety of methods may be used to find the solutions. The most direct method is to solve for α in the first equation ($\alpha = \bar{y} - \beta\bar{x}$) and substitute for α in the second equation to get β. The solutions are

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

Accordingly, we call

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

the **least squares estimates** of α and β, and

[5] The name has nothing to do with the normal distribution.

$$\hat{\alpha} = \bar{Y} - \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}\bar{x}$$
$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

the **least squares estimators** of α and β

You should note that the full assumptions (1) – (4) of the linear regression model are not needed to get least-squares estimates. It is only necessary that the model say that the value of y_i "should be" $\alpha + \beta x_i$. That's all! In particular, the normal assumption, homoscedasticity, and independence are not necessary to produce least-squares estimates. But if the full assumptions do apply, then the least-squares estimates are also maximum likelihood estimates. (Why?)

Basic properties (mean, variance, distribution) of the estimators can be derived most easily by writing them as linear combinations of the *Y*'s and applying the properties of linear combinations from Chapter 2. To do this, first observe that $\sum (x_i - \bar{x})(Y_i - \bar{Y}) = \sum (x_i - \bar{x})Y_i$. (Why?) Then

$$\hat{\alpha} = \bar{Y} - \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}\bar{x} = \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) Y_i = \sum v_i Y_i$$
$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) Y_i = \sum w_i Y_i$$

shows that the least-squares estimators can be written as linear combinations of the *Y*'s. The weights v_i and w_i are *not* random under the conditional regression model.

Since a linear combination of normally distributed RV's has a normal distribution (Chapter 2), therefore the least-squares estimators are normally distributed. To find the mean and variance of their normal distributions, apply Chapter 2 properties on the mean and variance of linear combinations:

$$E(\hat{\alpha}) = E\left(\sum v_i Y_i\right) = \sum v_i E(Y_i) = \sum \left\{ \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) (\alpha + \beta x_i) \right\} = \alpha - 0 + \beta\,\bar{x} - \beta\,\bar{x} = \alpha$$
$$E(\hat{\beta}) = E\left(\sum w_i Y_i\right) = \sum w_i E(Y_i) = \sum \left\{ \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right) (\alpha + \beta x_i) \right\} = 0 + \beta \sum \frac{(x_i - \bar{x})x_i}{\sum (x_i - \bar{x})^2} = \beta$$

and

$$Var(\hat{\alpha}) = \sum v_i^2 Var(Y_i) = \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 \sigma^2$$
$$Var(\hat{\beta}) = \sum w_i^2 Var(Y_i) = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 \sigma^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Thus, the least-squares estimators are unbiased, normally distributed, and have variances given by the above formulas.[6] The joint distribution of the least-squares estimators is bivariate normal

[6] Note that the means and variances cannot be calculated in this manner if the weights v_i and w_i are random – as they would be in the random *X* version of the linear regression model.

and has covariance that can be calculated using the formula for covariance of two linear combinations from Chapter 2:

$$Cov(\hat{\alpha},\hat{\beta}) = Cov\left(\sum v_i Y_i, \sum w_i Y_i\right) = \sum\sum v_i w_j Cov(Y_i, Y_j) =$$

$$\sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)\frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\sigma^2 = -\frac{\bar{x}}{\sum(x_i-\bar{x})^2}\sigma^2$$

Residuals.

The **error** that the model makes in saying that y_i "should be" $\alpha+\beta x_i$ is $y_i-\alpha-\beta x_i$. Since the true values of α and β are unknown, the true value of the error is unknown. But the error can be estimated by the **residual** $y_i-\hat{\alpha}-\hat{\beta}x_i$. If y_i is an "average" observation, it should be an "average" distance from the mean $\alpha+\beta x_i$ of its distribution. Thus, the magnitude of the single residual $y_i-\hat{\alpha}-\hat{\beta}x_i$ is an estimate of the "average" deviation σ, and $(y_i-\hat{\alpha}-\hat{\beta}x_i)^2$ is an estimate of σ^2. The average of the squared residuals should be a better estimate of σ^2 than a single squared residual. You can get an unbiased estimate of σ^2 by the average $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{n-2}$, and an unbiased estimator is $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{n-2}$. Note that the numerator of the estimate is the minimum value of the ESS. The denominator is not *n*. We divide by *n*-2 in order to get an unbiased estimator of σ^2. Unbiasedness follows from the fact that $\frac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{\sigma^2}$ has a χ^2 distribution[7] with *n*-2 degrees of freedom and so has a mean of *n*-2. The estimate $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{n-2}$ is called the **mean-squared error** (**MSE**), and its square root $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{n-2}}$ is called the **root mean-squared error** (**RMSE**). The RMSE is, in effect, the standard deviation of the residuals because the mean of the residuals is zero (why?). So the RMSE can be interpreted as an "average" amount by which the estimates $\hat{\alpha}+\hat{\beta}x_i$ miss their targets y_i, in the same way that the ordinary (sample) standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ can be interpreted as an "average" amount by which the estimate $\bar{y}$ misses

[7] It is tedious to show this, but it can be done with no more than the properties given in Chapter 2. To do this, start with the fact that $\frac{\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)^2}{\sigma^2}$ has a χ^2 distribution with *n* degrees of freedom. (Why?) Break this expression into parts that have χ^2 distributions with 1, 1, and n-2 degrees of freedom after adding and subtracting $\hat{\alpha}+\hat{\beta}x_i$ inside the parentheses.

its targets y_i. Thus, RMSE is a fundamental measure of how well the regression model fits the data.[8]

Hypothesis tests and confidence intervals for α and β

We noted above that

$$\hat{\alpha} \text{ has a normal distribution with mean } \alpha \text{ and variance } \sum\left(\frac{1}{n}-\frac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)^2\sigma^2$$

$$\hat{\beta} \text{ has a normal distribution with mean } \beta \text{ and variance } \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$$

These facts can be used in the usual way to make confidence intervals and test hypotheses about α and β, provided the value of σ^2 is known. But it is unrealistic to expect the value of σ^2 to be known. In the more realistic case that σ^2 is unknown, we can combine the normal distribution of $\hat{\alpha}$ and $\hat{\beta}$ with the chi-square distribution of $\dfrac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{\sigma^2}$ to make confidence intervals and test hypotheses about α and β with the *t* distribution:

The random variable $\left.\dfrac{\hat{\beta}-\beta}{\sqrt{\dfrac{\sigma^2}{\sum(x_i-\bar{x})^2}}}\middle/\sqrt{\dfrac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{\sigma^2(n-2)}}\right.$ is the ratio of a standard normal RV (why?) in the numerator to the square root of an independent[9] chi-square RV over its degrees of freedom. Therefore, this ratio, which equals $\dfrac{\hat{\beta}-\beta}{\dfrac{\hat{\sigma}}{\sqrt{\sum(x_i-\bar{x})^2}}}$, meets the requirements for having a *t* distribution with *n*-2 degrees of freedom.

Similarly, $\left.\dfrac{\hat{\alpha}-\alpha}{\sqrt{\sum\left(\dfrac{1}{n}-\dfrac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)^2\sigma^2}}\middle/\sqrt{\dfrac{\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{\sigma^2(n-2)}}\right. = \dfrac{\hat{\alpha}-\alpha}{\hat{\sigma}\sqrt{\sum\left(\dfrac{1}{n}-\dfrac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)^2}}$

meets the requirements for having a *t* distribution with *n*-2 degrees of freedom. These facts can be used in the usual way to make confidence intervals and test hypotheses about α and β when the value of σ^2 is unknown.

[8] The other fundamental measure of model fit is “R-square”.

[9] Independence is true, but takes some work to show.

In both of these facts, it appears that we have simply substituted the estimate $\hat{\sigma}$ for σ. That is the way it is usually explained in first courses: You don't know the value of σ, so you substitute a sample estimate for it. But we now see that there is more going on, although when the dust settles, it amounts to substitution.

What if the distribution of *Y* given *x* is not normal? If all other specifications for the regression model hold except (4) – normality – then the above facts are still approximately true, provided *n* is sufficiently large. This is a consequence of another version of the Central Limit Theorem. I won't discuss this in any detail, but I will mention why it is plausible: Recall from above that the least-squares estimators are linear combinations of $Y_1, Y_2, ..., Y_n$. If the weights in these linear combinations were all equal to 1/n, then the standardized least-squares estimators would have asymptotic normal distributions since they would be the mean $\bar{Y}$. But the weights are not all 1/n and the weights are not the same for every *Y*. Some *Y*'s get more weight than others. It turns out that the Central Limit Theorem still applies even if the weights are not the same, as long as the weights don't vary too much. The weights used in linear regression qualify.

The Gauss-Markov Theorem.

Under regression specifications (1) – (3) (all but normality), the least-squares estimators of α and β are optimal, in the sense that there are no other unbiased *linear* combinations of $Y_1, Y_2, ..., Y_n$ that have smaller variance. This is the Gauss-Markov Theorem. We say that the least-squares estimators are BLUE (Best Linear Unbiased Estimators). The Gauss-Markov Theorem does not completely rule out the possibility that there may be a nonlinear function of $Y_1, Y_2, ..., Y_n$ that is better. And it does not rule out the possibility that there might be a biased estimator that has smaller variance.

Inference on the conditional mean of *Y*

The conditional mean of *Y* given *X* = *x* (*x* is not necessarily a data value) is $E(Y \mid X = x) = \alpha + \beta x$. This is the regression equation itself. It is natural to estimate the conditional mean by plugging in the least-squares estimates for α and β:

$$\hat{E}(Y \mid X = x) = \hat{\alpha} + \hat{\beta} x$$

Is that a good idea? Yes. Since the least-squares estimators are unbiased, then $\hat{\alpha} + \hat{\beta} x$ is an unbiased estimator of $\alpha + \beta x$. (Why?) The *x* at which we are estimating the conditional mean is any value of *x* – not necessarily a data value. Since the least-squares estimators are linear combinations of $Y_1, Y_2, ..., Y_n$, then $\hat{\alpha} + \hat{\beta} x$ is a linear combination of $Y_1, Y_2, ..., Y_n$. And since $Y_1, Y_2, ..., Y_n$ given $x_1, x_2, ..., x_n$ are normal, then $\hat{\alpha} + \hat{\beta} x$ is a linear combination of normally distributed RVs and is therefore normal. Applying a formula from Chapter 2, the variance is

$$Var(\hat{\alpha} + \hat{\beta} x) = Var(\hat{\alpha}) + 2x Cov(\hat{\alpha}, \hat{\beta}) + x^2 Var(\hat{\beta})$$

Substituting the previously calculated values for the variances and covariance, we have

$$Var(\hat{\alpha} + \hat{\beta} x) = \sum \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 \sigma^2 - 2x \frac{\bar{x}}{\sum (x_i - \bar{x})^2} \sigma^2 + x^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

After algebraic simplification, this becomes

$$Var(\hat{\alpha}+\hat{\beta}x)=\sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

As usual, σ^2 may be estimated by the MSE, $\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{n-2}$, giving us a standard error for this estimate of $\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$.

So hypothesis tests and confidence intervals can be done using the RV $\frac{\hat{\alpha}+\hat{\beta}x-(\alpha+\beta x)}{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}$, which has a *t* distribution with *n*-2 degrees of freedom. (Why?) If the normal specification (4) in the model does not hold, then this statistic still has an approximate t_{n-2} distribution if *n* is sufficiently large.

Predicting an individual value of *Y* at a given *x*.

Suppose we have observed $y_1, y_2,..., y_n$ at corresponding $x_1, x_2,..., x_n$ and the regression model applies. Suppose that we have not yet observed the *Y* at a given *x*. We wish to estimate the outcome of *Y* at *x*. How to do this?

Since the regression model applies, then $Y=\alpha+\beta x+\varepsilon$. Now the error ε has mean zero, and because ε is independent of the RVs that we have observed so far, there is no way to tell whether ε will be positive or negative – although we could estimate its magnitude (how?). That is, the only part of *Y* that can be estimated unbiasedly is its mean $\alpha+\beta x$. We therefore estimate Y by $\hat{\alpha}+\hat{\beta}x$.

How well does this estimator do? The prediction error is $Y-(\hat{\alpha}+\hat{\beta}x)$. The prediction error is a RV that is a linear combination of normal variables (why?) and therefore has a normal distribution itself. The mean prediction error is zero (why?), and using a formula from Chapter 2, we can calculate the variance to be

$$Var(Y-(\hat{\alpha}+\hat{\beta}x))=Var(Y)-2Cov(Y,\hat{\alpha}+\hat{\beta}x)+Var(\hat{\alpha}+\hat{\beta}x)=Var(Y)+Var(\hat{\alpha}+\hat{\beta}x)=$$

$$\sigma^2+\sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)=\sigma^2\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

The reason that $Cov(Y,\hat{\alpha}+\hat{\beta}x)$ = 0 in this evaluation is that *Y* is a new observation, independent of the observations $Y_1, Y_2,..., Y_n$ made previously. Since $\hat{\alpha}+\hat{\beta}x$ is a function of the previous observations, then *Y* is independent of $\hat{\alpha}+\hat{\beta}x$ and so is uncorrelated with it.

As usual, σ^2 may be estimated by the MSE, $\hat{\sigma}^2=\frac{\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2}{n-2}$, giving us a standard error for this estimate of $\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$.

If you compare the variance formula for prediction error (immediately above) with the variance of the conditional mean estimator (a page earlier), you will see that they are the same, except that the prediction error variance includes an extra σ^2. The practical effect of this is that there is an irreducible error in predicting a new observation. No matter how many data you have collected, you can never reduce the standard error of prediction below σ. On the other hand, the variance of the conditional mean estimator can be driven as low as you please by collecting more data: As n increases, $\hat{\sigma}^2$ gets closer to σ^2 (why?), 1/n goes to zero, and $\dfrac{(x-\bar{x})^2}{\sum (x_i-\bar{x})^2}$ typically goes to zero (why?). So the variance of the conditional mean estimator $\sigma^2\left(\dfrac{1}{n}+\dfrac{(x-\bar{x})^2}{\sum (x_i-\bar{x})^2}\right)$ typically goes to zero.

R-square.

R-square is one of the two primary measures for assessing model fit. The other is RMSE. In my opinion, RMSE is neglected and should get equal emphasis along with R-square. But the research world is fixated on R-square. R-square provides a unit-less measure of fit, whereas the unit for RMSE is the same unit as for *Y*.

> A brief word in defense of RMSE: With RMSE interpreted as the "average" magnitude of the error in estimating *Y* by $\hat{\alpha}+\hat{\beta}x$, RMSE provides an understandable measure of typical error in units that have immediate physical meaning. A glance at RMSE tells the subject-matter expert immediately whether the typical error is large or small, especially in relation to the average *Y* being estimated. R-square lacks this immediacy since it is unit-less.

In brief, R-square assesses the improvement in ESS when knowledge of the *x*-values $x_1, x_2, ..., x_n$ is added to knowledge of the *y*-values $y_1, y_2, ..., y_n$. How is the assessment done? First, note that without the *x*-values, the best that can be done to estimate the *y*-values is to use the *y*-mean $\bar{y}$. Then the ESS is $\sum (y_i-\bar{y})^2$, which we now call **total sum of squares** (**TSS**). When the *x*-values are made available and linked with corresponding *y*'s, the ESS is $\sum_{i=1}^n (y_i-\hat{\alpha}-\hat{\beta}x_i)^2$. The change in ESS is $\sum (y_i-\bar{y})^2$ - $\sum_{i=1}^n (y_i-\hat{\alpha}-\hat{\beta}x_i)^2$, which we intuitively expect to be nonnegative (why?).

From the mathematical identity $\sum (y_i-\bar{y})^2$ = $\sum [y_i-(\hat{\alpha}+\hat{\beta}x_i)+(\hat{\alpha}+\hat{\beta}x_i)-\bar{y}]^2$ = $\sum [y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2+\sum [(\hat{\alpha}+\hat{\beta}x_i)-\bar{y}]^2$ [in the expansion of the square, the cross-product term is zero], we see that the change in ESS is $\sum (y_i-\bar{y})^2$ - $\sum_{i=1}^n (y_i-\hat{\alpha}-\hat{\beta}x_i)^2$ = $\sum [\hat{\alpha}+\hat{\beta}x_i-\bar{y}]^2$, which is nonnegative. So using the *x*'s improves the ESS. The improvement $\sum [\hat{\alpha}+\hat{\beta}x_i-\bar{y}]^2$ is called the **regression sum of squares** (**RSS**).

The percent improvement in ESS is RSS/TSS = $\sum [\hat{\alpha}+\hat{\beta}x_i-\bar{y}]^2/\sum (y_i-\bar{y})^2$, which is **R-square**. As a percent or proportion, R-square necessarily lies between 0 and 1. If the regression line fits the data perfectly, then $y_i=\hat{\alpha}+\hat{\beta}x_i$ for all *i*, so R-square = 1 (complete improvement results from using *x*). If the regression line is flat, then $\hat{\beta}=0$ and $\hat{\alpha}=\bar{y}$ (to see this, check the formula for $\hat{\alpha}$), so R-square = 0 (no improvement from using *x*).

By substituting the values $\hat{\alpha} = \bar{y} - \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \bar{x}$ and $\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ into R-square = $\sum [\hat{\alpha} + \hat{\beta} x_i - \bar{y}]^2 / \sum (y_i - \bar{y})^2$ and simplifying, you can also show that R-square = $\frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$, which is the square of the correlation coefficient between the *x*'s and *y*'s. This explains the origin of the name *R-square*, for *r* is the symbol commonly used for the sample correlation coefficient.

Testing the correlation coefficient.

In one-predictor regression, $\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ = $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}}$ = $r \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}}$ = $r \frac{s_y}{s_x}$. As long as there are at least two distinct *y*'s and *x*'s, this shows that the sample slope = 0 if and only if the correlation coefficient = 0. Therefore, the *t*-test previously given for the sample slope can be used to test $H_0 : \rho = 0$, where ρ represents the population correlation coefficient.

Another approach to testing the correlation coefficient uses an *F*-test. But this approach turns out to be essentially the same as testing the slope: By algebra, it can be shown that when $\rho = 0$ (equivalently, $\beta = 0$), the square of the *t*-statistic for testing $H_0 : \beta = 0$ is $(\hat{\beta} - 0)^2 \Big/ \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$ = $\frac{\sum [\hat{\alpha} + \hat{\beta} x_i - \bar{y}]^2 / \sigma^2}{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 / [\sigma^2 (n-2)]}$ = $\frac{RSS / \sigma^2}{ESS / [\sigma^2 (n-2)]}$ = $\frac{r^2 (n-2)}{1 - r^2}$ is the ratio of two independent χ^2 RVs over their degrees of freedom (1 d.f. in the numerator and *n*-2 d.f. in the denominator). Therefore, the square of this *t*-statistic qualifies as having an *F* distribution with 1 and *n*-2 degrees of freedom. So the F-test can be used to test $H_0 : \rho = 0$.

Finally, I mention Fisher's *z*-transformation. In 1915, Sir Ronald Fisher discovered a transformation of *r* that has a distribution that is approximately normal and a variance that is approximately constant. For tests and confidence intervals, Fisher's *z* is preferred over *r* itself because *z* is much closer to normality for small samples than *r* is.[10]

$$\text{Fisher's } z = \operatorname{arctan} h(r) = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right).$$

Its approximate variance is $\frac{1}{n-3}$. If *r* = 0, then *z* = 0. As $r \to 1,\ z \to \infty$. As $r \to -1,\ z \to -\infty$.

[10] *r* takes longer to converge to normality than most statistics – you may need a large *n* in order to be "sufficiently large" to assert normality of *r*.