



## Linear Models (continued)

### Model Selection

#### *Introduction*

Most of our previous discussion has focused on the case where we have a data set and only one fitted model. Up until this point, we have discussed, for one model with a particular set of explanatory variables, what fitting procedure should be used (e.g., least squares, generalized least squares, robust estimation, resistant estimation), how the estimated model coefficients should be interpreted and tested for significance, how goodness of fit can be evaluated for the model, how to detect problems with the assumptions made when fitting the model, and how to detect outliers. The main issue that remains to discuss is, if we have 2 or more models, how do we decide which one is the best for the data?

#### *Different potential models*

To begin our discussion of this topic, how might the models we are comparing differ? Well, first, two linear models might differ because the response variable takes a different form in each: for instance, in our Fuel Consumption example, the response variable in model *A* might be the outcome variable in its original form (Fuel Consumption in City), whereas the response variable in model *B* might be a transformation of the original outcome variable (e.g.,  $1/(\text{Fuel Consumption in City})$ ) might be a good response variable to use because Fuel Consumption can be thought as mileage/fuel or as fuel/mileage). Second, two linear models might differ because they include different explanatory variables from among the set of *potential candidate variables*. Note that the set of potential candidate variables includes the available *original candidate variables* (the design and background variables in their original form) and, additionally, any *derived candidate variables* (any transformation of one or more of the original candidate variables, such as a log or power transformation of one original variable or a product of two or more original variables). For instance, if, in our Fuel Consumption example, model *A* contains two explanatory variables, Engine Size and Weight, model *B* might differ because it also contains the explanatory variable Type (an additional original variable) or because it contains the derived variable  $(\text{Weight of Car})^2$  or  $(\text{Engine Size}) * (\text{Weight})$  or both.

As noted in our discussion of model checking, transforming the response variable for a particular linear model is one possible approach if problems of non-

normality, heteroscedasticity, or non-linearity are detected for the model. More specifically, one will often compare models with different transformations of the outcome variable (but the same explanatory variables) in order to find the *best transformation* (the transformation for which the assumptions of normality, constant variance, and linearity appear to be the most valid). For instance, for a particular set of explanatory variables, one might use the Box-Cox procedure to choose the best (power) transformation of the outcome variable. However, in the following discussion of model selection, we will ignore this first type of model difference and assume that all linear models being compared have the same response variable (i.e., the same form of the same outcome variable): we discuss only how to choose the *best set* of explanatory variables for that particular response variable. The model with the best set of explanatory variables (for the particular response form we've decided to use) will be termed *the final model*. It is important to be aware that, in some circumstances, we may end up with several final models rather than just one final model. For instance, in some situations, the researcher might want to report both a simple model and a complex model. In other instances, the researcher may report two models, each containing one member of a pair of highly correlated explanatory variables, because he is not certain which of these variables is affecting the response variable (and doesn't want to include both because they're highly correlated). It is also important to be aware that, in actual data analysis, choosing a transformation for the response variable should perhaps occur in conjunction with choosing the explanatory variables and, thus, candidates for the final model may have different forms of the outcome variable in addition to different sets of explanatory variables.

Before giving an overview of what it means for a particular set of explanatory variables to be the best, we should note that it is not only our view of which variables affect the response that is altered by the choice of explanatory variables in the final model. Our view of how the explanatory variables affect the response is also altered by the choice of explanatory variables. This is the case because the estimated value of a regression coefficient for a particular explanatory variable (which we typically use to make quantitative statements about how that variable affects the response) will typically differ depending on what other explanatory variables are in the model. There are only two situations in which the estimated regression coefficient for an explanatory variable does not change when a new explanatory variable is added to the model: when the new explanatory variable has no effect on the response (then why add it?) and when the new explanatory variable is completely uncorrelated with those explanatory variables already in the model.

#### *Overview of model (or explanatory variable) selection*

However, as we will see below, deciding which set of explanatory variables is the "best" is far from a simple task. First of all, the definition of best will vary depending on the situation, especially on the analyst's goals and reasons for fitting a linear model to the data (e.g., prediction vs. explanation). Further, even in a particular situation, the definition of "best" may be far from clear. What's more, even once a definition of best is arrived at, it may be hard to decide upon a means of quantifying how well models adhere to this definition.

In general, though, we would want to select the set of explanatory variables for which the resulting model (a.) appears to satisfy the assumptions of linear modelling or, at least, is closest to satisfying them and (b.) fits the data well and is as simple as possible.

Consideration (a.) is often ignored, as is the case when analysts pick models by merely comparing their respective values for a goodness of fit measure. However, since the conclusions drawn from a particular model can be highly erroneous if the assumptions of linear modelling are not satisfied, it is important to keep this consideration in mind. Thus, even though one model might fit the data better, we may prefer another, less well fitting model because it appears that the linear model assumptions are more valid for the second model, as indicated by various model checks. When faced with a trade-off between a better fitting model and a model that better satisfies the assumptions, an analyst's choice will depend on the particular goals and reasons why he is trying to model the data set. For instance, in some circumstances, it may be essential that the linear model assumptions are satisfied, and, thus, the analyst would choose the best set of explanatory variables (or, identically, the best model) from among only those models that satisfy these assumptions. In the following discussion of model selection, we will assume that we are considering only linear models that appear to satisfy the assumptions equally well (which we can ascertain by using the diagnostic techniques that were introduced in our previous discussion of Model Checking). Thus, we will not worry about consideration (a.).

Then, our discussion of model selection will focus on consideration (b.). Unfortunately, though, deciding which linear model fits the data best is far from straightforward. First, as we have previously discussed, it is unclear how to measure, for a particular linear model, how well it summarises the relationship between  $Y$  and  $X$ . In addition, even if there were a single accepted way to measure this, it would still not be obvious that the model with the higher value for this measure should be selected because of the often competing goal of simplicity or parsimony. Usually, we would prefer a simpler model (i.e., a model with fewer explanatory variables) to a more complicated one, particularly when explanation is the goal: this preference is backed by both philosophical reasons (e.g., simpler models are more elegant) and statistical reasons (complex models may be too specific to the data set that was used to fit them). However, in general, more complicated models tend to "fit the data" better as measured by typical goodness of fit criteria (e.g.,  $R^2$ ). Thus, we often face a trade-off between goodness-of-fit and simplicity when choosing between models. Of course, what place along the goodness of fit vs. simplicity spectrum we are comfortable with will depend on the particular situation.

Before addressing the issue of comparing how well models with different sets of explanatory variables meet consideration (b.), we will take a brief detour to discuss how to interpret the coefficients of derived explanatory variables.

## *Models with derived explanatory variables*

In general, we may want to measure the effect of a particular variable of interest on the response variable, which, as we have discussed previously, can be quantified using the regression coefficients. However, when the variable of interest is included among the linear model's explanatory variables not just as itself (as an original variable) but also as a derived variable, then the interpretation of its regression coefficients become somewhat more complicated.

We have already discussed how to interpret the coefficient of a continuous explanatory variable or a categorical explanatory variable treated using the factor approach with a treatment parameterisation. (We review these interpretations below). However, these coefficient interpretations are only valid when the corresponding explanatory variable appears in no other terms in the linear model. In some instances, such as when power terms or product terms are included in hierarchical models, this will not be the case; thus, below, we discuss the interpretation of coefficients in these instances. Further, our previous interpretation for a continuous explanatory variable coefficient was in terms of the effect of changes in that explanatory variable. Often, however, that explanatory variable is derived from an original variable, and we might prefer the interpretation to be in terms of changes in the original variable. Below, we discuss the appropriate interpretation for one instance of this phenomenon – when the explanatory variable corresponding to the coefficient in question is a logarithmic transformation of the original variable.

Before proceeding to discuss interpretations of coefficients, let us note that we might want to include higher order powers of an explanatory variable because we suspect that some curvature (i.e., non-linearity) is present in the relationship between that explanatory variable and the response variable. Similarly, we might want to include a product or interaction term for two or more explanatory variables because we suspect that the way in which each explanatory variable affects the response variable depends on the value of the other explanatory variables. In other words, choosing to include an interaction term for multiple explanatory variables is equivalent to stating that the collective effect of these explanatory variables on the response variable is more than the sum of their individual effects. These suspicions that a higher order and/or product term should be included may come from theoretical knowledge of the phenomenon being modelled or, alternatively, from looking at various diagnostic plots.

Finally, note that all of the following interpretations are “conditional on the other explanatory variables being the same” or “for units with approximately the same values of the other explanatory variables.”

### *Continuous explanatory variable*

#### **Appearing in only one term**

First, let us address the situation where the variable of interest is continuous and appears as the explanatory variable in one term in the linear model equation. The coefficient for that term is interpreted as the additive amount by which the mean response increases or decreases when the explanatory variable increases by 1 unit.

Note that this increase (decrease) in mean response is the same no matter which level of the explanatory variable we start at. For instance, in our Fuel Consumption example, the coefficient of Weight is the additive increase/decrease in mean Fuel Consumption that occurs when Weight increases by one pound, for cars of the same Engine volume, Cylinders, Type, and Drive Type.

However, what if the explanatory variable is the logarithm of the original variable in which we are interested? If that is the case and we want to interpret the coefficient in terms of changes in the original variable, then we would say that the explanatory variable's coefficient multiplied by  $\log(2)$  is the additive increase (decrease) in mean response that occurs when the original variable is doubled. For instance, what if, in our Fuel Consumption example, our model includes  $\log(\text{Weight})$  instead of Weight as an explanatory variable. Well, the way we interpret the coefficient of  $\log(\text{Weight})$  is the additive increase (decrease) in mean Fuel Consumption that occurs when  $\log(\text{Weight})$  increases by one log-pound, for cars with the same Engine, etc . . . . However, if we wanted to interpret the coefficient in terms of changes in Weight rather than  $\log(\text{Weight})$ , then we could say that the coefficient of  $\log(\text{Weight})$  times  $\log(2)$  is the amount added to (subtracted from) mean Fuel Consumption when Weight doubles, for cars with the same Engine, etc . . . .

#### Also appearing in a squared term

Suppose that our continuous variable of interest,  $X_j$ , is included in the linear model as itself and as a squared term. In this case, we would interpret together all the coefficients of terms that include  $X_j$ . More specifically, suppose the model includes the terms  $\cdots + \beta_j X_j + \beta_{j,sq} X_j^2 + \cdots$ . Then, the two coefficients pertaining to  $X_j$  would be interpreted in the following manner: when  $X_j$  starts at value  $x_j$  and increases by 1 unit, the additive increase (decrease) in mean response is  $\beta_j + \beta_{j,sq}[2x_j + 1]$ . Note that, in this case, the effect that changes in  $X_j$  have on the mean response depends on what value  $X_j$  started at. For instance, suppose our Fuel Consumption model includes the explanatory variables Weight and  $(\text{Weight})^2$ . Well, we would interpret their coefficients, together, by saying: if Weight starts at 4000 pounds, then increasing Weight by 1 pound increases (decreases) mean Fuel Consumption by an amount equal to the coefficient of Weight plus 8001 times the coefficient of  $(\text{Weight})^2$ , for cars with the same Engine, etc . . . .

#### Also appearing in a product term (with another continuous variable)

Suppose that our continuous variable of interest,  $X_j$ , is included in the linear model as itself and in a product term with another continuous variable. In this case, we would interpret together the two coefficients pertaining to  $X_j$ . More specifically, suppose the model includes the terms  $\cdots + \beta_j X_j + \beta_{j,s} X_j X_s + \cdots$ , where  $X_s$  is another continuous variable. Then, we interpret the coefficients pertaining to  $X_j$  in the following manner: when  $X_j$  increases by one unit and  $X_s$  is at value  $x_s$ , the additive increase (decrease) in mean response is  $\beta_j + \beta_{j,s} x_s$ . Note that, in this case, the effect that changes in  $X_j$  have on the mean response depends on what value  $X_s$ , the other continuous variable in the product term, starts at. In fact, this is exactly what it means

when we say that there is an interaction between two variables in the way they affect a response variable: the way in which one variable affects the response variable depends on the value of the other variable. As an example, suppose our Fuel Consumption model includes the explanatory variables Weight and Weight\*Engine Size. Well, we would interpret their coefficients, together, by saying: if Engine Size starts at 10 units of volume, then increasing Weight by 1 pound adds (subtracts) an amount equal to the coefficient of Weight plus 10 times the coefficient of Weight\*Engine Size to mean Fuel Consumption, for cars with the same Engine, etc . . . .

### *Categorical explanatory variable*

#### **Appearing alone**

If our categorical variable of interest is treated using the factor approach with a treatment parameterisation, then the coefficient value for each (non-baseline) level of the categorical variable is the difference between the mean response at that level and the mean response at the baseline level.

#### **Also appearing in a product term**

Suppose that our categorical variable of interest,  $X_j$ , is included in the linear model as itself and in an interaction with a continuous variable. In this case, assuming that the factor approach is used for the categorical variable, then the linear model would contain two terms (one regular term, one interaction term) for each non-baseline level of our categorical variable: for example, for level  $k$  of the categorical variable, the linear equation would include the terms  $\cdots + \beta_{j,k} X_{j,k} + \beta_{j,k,s} X_{j,k} X_s + \cdots$ , where  $X_s$  is a continuous variable and, if a treatment parameterisation is used,  $X_{j,k}$  is an indicator variable for level  $k$  of variable  $j$ . Then, the two coefficients for level  $k$  of the categorical variable are interpreted as follows: if variable  $X_s$  has the value  $x_s$ , then  $\beta_{j,k} + \beta_{j,k,s} x_s$  is the difference between the mean response at level  $k$  and the mean response at the baseline level. Note that, again, the effect on the mean response of changing from the baseline level to level  $k$  of variable  $X_j$  depends on what value  $X_s$ , the other variable in the product term, is at. As an example, suppose that our Fuel Consumption model contains both Type and an interaction between Type and Weight; further, suppose that Type is treated using the treatment version of the factor approach. Then, assuming that "compact" is the baseline level in the treatment parameterisation, the coefficient of the sporty type indicator plus 4000 times the coefficient of the interaction between the sporty type indicator and Weight would be the difference in mean Fuel Consumption between sporty and compact cars for cars that weigh 4000 pounds and have the same engines, number of cylinders, etc. . . .

Now, suppose that our categorical variable of interest,  $X_j$ , is included in the linear model both as itself and in an interaction with another categorical variable. In this case, assuming that we used the treatment version of the factor approach for both categorical variables and that the variables have  $K$  and  $T$  levels, respectively, then our linear model would contain  $(K-1)(T-1)$  interaction terms (one for every combination of non-baseline levels),  $(K-1)$  regular terms for the first variable, and  $(T-1)$  regular terms for the second variable. The coefficient of the interaction term pertaining to level  $k$  of variable 1 and level  $t$  of variable 2 would then be interpreted as the extra additive

change in the mean response that occurs when levels  $k$  and  $t$  occur together (compared to adding the individual effect on the response of level  $k$  occurring to the individual effect of level  $t$  occurring). Although this interpretation may currently be a bit unclear, it will hopefully become less opaque later when we discuss ANOVA. As an example, suppose that our Fuel Consumption model contains Type and Cylinder as explanatory variables and also an interaction between the two. Further, suppose that both Type and Cylinder are treated using the treatment version of the factor approach (with “compact” and “three cylinders” as the respective baseline levels). Then, the coefficient for the interaction term between, say, sporty and four cylinders would be interpreted as the extra change in mean Fuel Consumption (relative to compact, three cylinder cars) that occurs for sporty and four cylinder cars compared to the change for sporty cars (relative to compact ones) plus the change for four cylinder cars (relative to three cylinder ones), provided that the cars have the same Weight, Engine, etc. . . .

### *Choosing between models (or between sets of explanatory variables)*

As stated previously, we will assume that we are only seeking to compare models with the same response variables, which reduces the model selection issue to the task of choosing the best set of explanatory variables. What’s more, we will assume that all models being considered appear to satisfy the linear model assumptions equally well, which means that we will choose a set of explanatory variables by comparing how well the resulting models “fit the data” and adhere to the principle of simplicity.

We discussed above that it can be difficult to decide how to measure goodness of fit for linear models. Further, we noted that even once we decide on a way to do so, it is unclear that we should just choose what is the best fitting model according to some strict measure of goodness of fit because of the competing goal of simplicity. For instance, suppose we use  $R^2$  to assess goodness of fit for competing models. Well, it is a mathematical fact that  $R^2$  will increase (or at least stay the same) whenever we add another explanatory variable into the model (while keeping all already included explanatory variables in the model); thus,  $R^2$  favours more complex models (i.e., models with more explanatory variables). If we also value simplicity in addition to goodness of fit, then we should not just select the model with the highest  $R^2$  value but should pick one that is reasonably simple but still has a high enough  $R^2$  value. Alternatively, we might select a model using adjusted- $R^2$  because it takes model complexity into account: it is roughly equal to  $R^2$  minus a penalty for increasing complexity, so that adjusted- $R^2$  doesn’t necessarily increase whenever we add an explanatory variable and thus favour the largest model. This example suggests that, instead of basing our model selection on a strict goodness of fit criterion that only rewards goodness of fit, such as  $R^2$ , we will prefer to use a *complexity-adjusted* goodness of fit criterion, such as the  $F$ -statistic, adjusted- $R^2$ ,  $C_p$ , AIC, or BIC, all of which reward goodness of fit and penalise complexity (or, identically, reward simplicity).

When we use one of these complexity-adjusted goodness of fit criteria to compare how well different models (with different explanatory variables) fit the data, taking simplicity into account, we do not necessarily use the criterion of choice on the

same data set that we used to fit our models (i.e., estimate their parameters). If we attempt to measure complexity-adjusted goodness of fit on the data set we used to fit a particular model, we will probably end up with an overly optimistic picture of how well the model summarises the relationship between  $X$  and  $Y$ ; because the model's parameters were estimated using this data set, we should expect that it will be a pretty good fit to that data set and that it might not fit another data set equally well. In fact, if we base our model selection on comparisons of a complexity-adjusted goodness of fit criterion for the data set used to fit the models, we will likely end up choosing a more complex model that is too specific to that particular data set. (This phenomenon is referred to as *overfitting*.) Ideally, then, we would want to compare the complexity-adjusted goodness of fit of different models using an entirely new data set, saved for this purpose and referred to as a *validation data set*. Unfortunately, though, such a data set is not often available to us. For this reason, we may instead use a *cross-validation* technique. These techniques entail breaking the only data set we have into two parts, using one part to fit the different models' parameters and the other part to measure the models' complexity-adjusted goodness of fit; this is typically done repeatedly and a model is chosen based on the combined assessments of the model's complexity-adjusted goodness of fit from all of the repetitions.

Before proceeding to discuss various complexity-adjusted goodness of fit criteria, we should delineate the difference between nested models and non-nested models because some frequently used model comparison techniques are only appropriate for comparing nested models. *Nested models* are models that are the same with the only difference being that certain parameters in the outer model are set to zero in the inner model. What does this definition mean with regards to different linear models (i.e., models with different explanatory variables) for the same response variable? Well, as a more concrete definition, model  $A$  is nested within model  $B$  if model  $A$  contains a subset of the explanatory variables in model  $B$ ; we will refer to model  $A$  as the *inner model* and model  $B$  as the *outer model*. In our Fuel Consumption example, a model with explanatory variables Weight and Cylinders is not nested within a model with explanatory variables Weight, Engine Size, and Type because  $\{\text{Weight, Cylinders}\}$  is not a subset of  $\{\text{Weight, Engine Size, Type}\}$ . However, a model with explanatory variables Weight and Cylinders is nested within a model with explanatory variables Weight, Cylinders, Engine Size, and Type because  $\{\text{Weight, Cylinders}\}$  is a subset of  $\{\text{Weight, Cylinders, Engine Size, Type}\}$ . To see how our second, more concrete, definition accords with our first definition of nesting, note that we can view the inner model as the same as the outer model except for the fact that the regression coefficients for any explanatory variables in the outer model but not in the inner model are set to zero. For instance, in our preceding example of nested models, the first (inner) model is the same as the second (outer) model except for the fact that the regression coefficients of Engine Size and Type are set to zero in the inner model.

### *Nested model comparison*

Deciding between two nested models amounts to deciding whether those explanatory variables in the outer model but not in the inner model (for convenience, we will refer to these as the *differing explanatory variables*) should be included; should

they be added to the inner model or should they be deleted from the outer model? In other words, do the differing explanatory variables significantly improve our explanation/prediction of  $Y$  over using just the explanatory variables in the inner model? For instance, in our previous example, comparing the inner and outer models is asking whether Engine Size and Type add any value over Weight and Cylinders in terms of explaining/predicting Fuel Consumption.

We might want to put the preceding question into quantitative terms in order to make it possible for us to mathematically/statistically answer it. As can be easily seen, not having the differing explanatory variables in the model is the same as having the coefficients of all differing explanatory variables be zero in the outer model. Thus, we can decide whether the differing explanatory variables should be in the model—whether they add any value over the explanatory variables already in the inner model—by testing the hypothesis that the coefficients of all differing explanatory variables are zero (in the outer model).

### **Nested models differing by one variable**

In some model comparisons, there may be only one differing explanatory variable between two nested models, so that comparing them amounts to testing whether this one variable should be in the model or, identically, whether its coefficient in the outer model is zero. As we have discussed previously, we can test the hypothesis that the regression coefficient for a particular explanatory variable is zero by looking at the  $t$ -value and  $p$ -value that accompany that coefficient in the typical regression output. Thus, we may decide that the differing explanatory variable should not be in the model if, when we look at the regression output for the outer model (which includes the variable), that variable is not significant at some particular level (i.e., its  $p$ -value is larger than some cutoff, such as  $0.05$ , or, identically, its  $t$ -value is between  $\pm 2$ ). Interestingly, using this approach with a  $t$ -value cut-off of  $\pm 1$  is the same as choosing between the inner and outer model by picking the one with the higher adjusted- $R^2$  value: this is the case because adjusted- $R^2$  only increases when an explanatory variable is added to the model if the  $t$ -value of that entering explanatory variable is greater than  $1$  or smaller than  $-1$ . (Contrast this to  $R^2$ , which increases (or at least stays the same) whenever an explanatory variable is added to the model, regardless of the  $t$ -value of that entering explanatory variable.)

#### *A brief digression on partial regression plots*

In the above, we are deciding whether one particular differing explanatory variable should be left out of the model or, identically, deciding whether it should be included in the model. However, to arrive at this stage, we need to have an idea that this particular variable might add some value to our linear model. But how do we get an idea of which potential candidate variables should perhaps be added to the existing model? Well, a particularly convenient and informative way of doing so is to look at a **partial regression plot** for every potential candidate variable not already in the model; if there is a trend, as opposed to just a random scatter of points, in the plot for a particular candidate variable, then we might suspect that the variable does have an influence on  $Y$  (even alongside those explanatory variables that are already in the model) and that the variable should be added to the model. This plot essentially shows us what least squares fitting for our multiple regression model is seeing when it adds

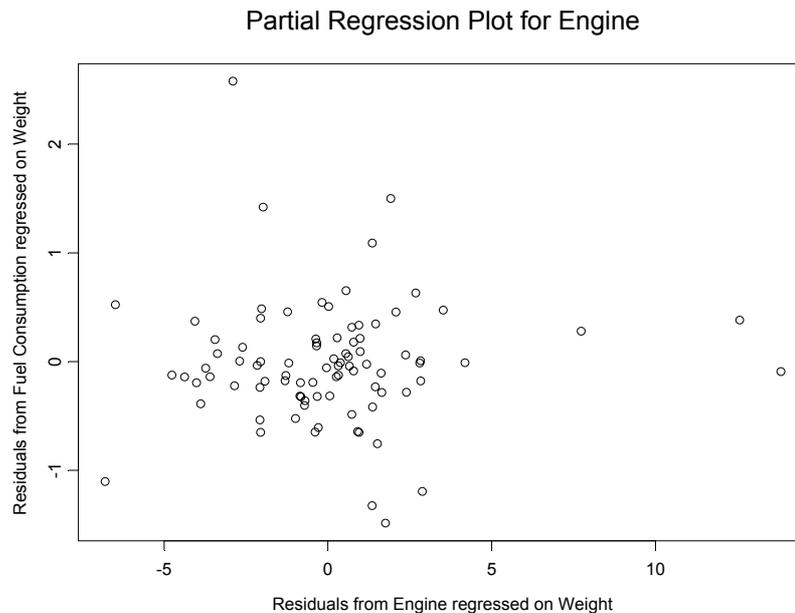
the particular candidate variable as an explanatory variable: the slope of the least squares line through the scatter of points in the partial regression plot is exactly the same as the least squares estimate of the regression coefficient for the candidate variable when it is added to the already existing model, and this estimated coefficient will be more significant the more closely the points adhere to the line in the plot. However, the partial regression plot contains more information than we would get by merely adding the candidate variable into the model, fitting the model to the data, and then examining its estimated coefficient and  $t$ -value/ $p$ -value. This is the case because the shape of trend we observe in the partial regression plot can give us an idea of what form of the candidate variable should be added to the model: a linear trend in the points might indicate that just the variable itself should be added to the model, whereas a curved trend in the points might indicate that we should also consider including higher power transformations of that variable in the model.

Technically speaking, the partial regression plot for a particular candidate explanatory variable is formed by (a.) finding the raw residuals for the existing linear model where the response variable,  $Y$ , is regressed on the set of already included explanatory variables, (b.) finding the raw residuals for another linear model in which the candidate explanatory variable is regressed on the already included explanatory variables (here, the candidate explanatory variable is treated as the response variable), and (c.) plotting the first set of residuals against the second set. However, the details of making a partial residual plot are unimportant: all you have to remember is that the presence of a trend in the points in this plot indicates that you might consider including the particular candidate variable in the model as an explanatory variable, and that the shape of the trend can tell you what forms of the variable you might include.

Partial regression plots are useful for investigating if we should perhaps include a specific potential candidate variable in our existing model, whether this specific variable is a new variable entirely unrelated to those explanatory variables already in the model or, instead, a variable derived from one of the variables already in the model. In this second case, we must know specifically what form (i.e., squared, logged) of the already included variable we might want to add in order to be able to use a partial regression plot. However, in some instances, we may just want to get an idea of whether we should add additional transformed versions of an already included variable and may not have any idea of what specific transformation should be used. Well, for this purpose, it can be very useful to plot the residuals from the existing model (where  $Y$  is regressed on the already included explanatory variables) against each of the already included explanatory variables (for explanatory variable  $j$ , we will call this a plot of  $e_i$  vs.  $X_{j,i}$ ). If we see a trend in the residuals in the plot for (already included) explanatory variable  $j$ , then we might want to consider including various transformations of  $X_j$  in our model. Essentially, a trend is telling us that the unexplained part of  $Y$ , as measured by the residuals, still depends on  $X_j$ ; in other words,  $X_j$  (or some form of it) still has some explanation power for  $Y$  left. (Note that a trend in the magnitude, or absolute value, of the residuals can be a sign of heteroscedasticity, indicating that the variance of the errors may not stay constant as  $X_j$  changes.)

**Example #1 (continued)**

As an example, consider again our Fuel Consumption study. Suppose that we already have a model in which Fuel Consumption is explained by Weight. Further, suppose we're trying to see whether we might want to add Engine Size as an explanatory variable and, also, whether we should include just Engine Size itself or possibly some transformations of it. Well, these questions can be answered by looking at the following partial regression plot for Engine Size given a model that already includes Weight:



*Figure 1: Should Engine Size be added as an explanatory variable to a model where Weight explains Fuel Consumption?*

In the above partial regression plot, the scatter of points appears fairly random; there does not seem to be a trend. Thus, we suspect that Engine Size does not significantly improve our explanation of Fuel Consumption once Weight is already in the model.

However, what if we disregard this evidence and decide to try adding Engine Size into the model anyway? Well, to see whether Engine Size should be kept in the model, we can examine its t-value/p-value in the computer output for the new model:

Coefficients:				
	Value	Std. Error	t-value	p-value
(Intercept)	47.1388	2.0106	23.4453	0.0000
Weight	-0.0081	0.0010	-8.0237	0.0000
Engine Size	0.0476	0.5743	0.0829	0.9341

As we can see from the last row in the above table, the suspicions raised by the partial regression plot are confirmed. The estimated coefficient for Engine Size

(0.04) is very close to zero (relative to its error), which accords with the fact that the least squares line through the scatter of points above would be basically horizontal, having zero slope. Further, the coefficient of Engine Size is highly insignificant ( $t$ -value far below 2 or even 1, and  $p$ -value about 0.90), leading us to conclude that Engine Size doesn't add much value once Weight is already in the model ■

### Nested models differing by more than one variable

In other comparisons of two nested models, there may be more than one differing explanatory variable. In these cases, we look at whether this whole *set* of explanatory variables should be added to the inner model (or deleted from the outer model), whether, taken together, they add value to our explanation/prediction of  $Y$  alongside those explanatory variables that are already included in the inner model. In statistical terms, we test the hypothesis that the regression coefficients for all of these differing variables should be zero in the outer model. For instance, in our Fuel Consumption example, we might want to test whether the coefficients of Type and Engine Size are both zero if Cylinders and Weight are already included in the model. Another frequent instance of testing whether a group of explanatory variables should be in the model occurs when we want to test whether a categorical variable, treated using a factor approach, should be included in the model. Remembering back, we recall that treating a categorical variable using the factor approach results in the inclusion of  $k - 1$  explanatory variables (of the indicator variety, if a treatment parameterisation is used) in the model. Thus, assessing whether the categorical variable is a significant predictor of  $Y$  amounts to assessing whether these  $k - 1$  indicator variables should be included in the model, or, to testing whether their coefficients are all zero in the outer model that includes them. (Here, note that the  $t$ -value/ $p$ -value of an individual indicator for one level of the categorical variable can be used to test whether that particular indicator should be in the model; if we don't reject the null hypothesis that the indicator's coefficient is zero, then we might consider dropping that level indicator variable from our model, which can be thought of as combining that level with the baseline level. However, finding a large  $p$ -value for one indicator variable should not lead us to remove the entire categorical variable from the model; we should only do so if we find large  $p$ -values for all of the variable's  $k - 1$  level indicators.)

Whether a *set* of explanatory variables should be included in the model can be tested using the *F-test for the joint significance of several terms*. Essentially, this test compares the Explained Sum of Square (ESS) of the outer model to the ESS of the inner model (with a standardisation factor that makes the comparison unit-free since the size of ESS will depend on the units in which  $Y$  was measured). Essentially, the more value that is added by the differing explanatory variables, the larger the outer model's ESS will be relative to the inner model's and, thus, the greater the difference in their two ESSs. However, because we know that  $R^2$  always increases whenever an explanatory variable is added, it is easy to see that ESS must increase too since it is just the

numerator of  $R^2$  (and the denominator stays the same). Thus, the ESS for the outer model will always be greater than the ESS for the inner model. To correct for the fact that ESS favours complexity, the  $F$ -test incorporates a penalty for complexity: specifically, the difference in ESS is divided by the number of differing explanatory variables. As a result, the same size increase in ESS is seen as less evidence in favour of the outer model when it occurs as a result of adding a larger number of explanatory variables. Generally speaking, if the complexity-adjusted (and standardised) difference in ESSs for the outer and inner models is large, then the  $p$ -value for the  $F$ -test will be small, leading one to reject the null hypothesis that the coefficients of the differing explanatory variables are all zero and that they should be left out of the model. Note that, in the above  $F$ -test, the null hypothesis is that all differing explanatory variables have zero coefficients; the alternative hypothesis is that at least one differing explanatory variable has a non-zero coefficient.

### **Example #1 (continued)**

Suppose that, in our Fuel Consumption example, we want to see whether the categorical variable Type, treated using the factor approach, should be included in the existing model with Weight. Well, including Type in the model using a factor approach with the treatment parameterisation means that five indicator explanatory variables are included in the model, and, thus, we need to test whether all five of their coefficients are zero. The output for this new model is:

Coefficients:				
	Value	Std. Error	t value	p-value
(Intercept)	44.2453	3.1911	13.8652	0.0000
Weight	-0.0074	0.0011	-6.9341	0.0000
TypeLarge	1.4187	1.3987	1.0143	0.3133
TypeMidsize	0.4178	1.0759	0.3884	0.6987
TypeSmall	2.6982	1.1523	2.3416	0.0215
TypeSporty	-1.0383	1.0533	-0.9858	0.3270
TypeVan	1.0531	1.5436	0.6823	0.4969

Note that “compact” is the baseline level in the above. Trying to decide whether to include Type based on the  $p$ -values for the coefficients of the five individual level indicators is a bit confusing since one is significant (e.g., Small), indicating that we may want to include Type, whereas the rest are fairly large, indicating that we may not want to include Type. Thus, we look at the results of the  $F$ -test for the joint significance of these five terms: the  $p$ -value for the  $F$ -statistic is 0.013, suggesting that we should reject the null hypothesis that all five coefficients are zero and suggesting that Type does indeed add some value to our model, even with Weight already in the model. Looking at the output above, we suspect that most of this value is coming from the differing Fuel Consumption performance of Small cars ■

Before proceeding, we should note two special cases of the preceding  $F$ -test. The first special case occurs when the differing explanatory variables include all of the explanatory variables in the outer model, which means that the inner model just has an intercept term and no explanatory variables. In this case, the  $F$ -test discussed above is

the same as the previously discussed *overall F-test*. The other special case occurs when there is only one differing explanatory variable; in this case, performing the above *F-test* is the same as looking at the *t-value/p-value* of that one explanatory variable in the outer model.

### *General model comparison using various criteria*

In general, regardless of whether the models we are considering are nested or not, we might choose between them based on their values for one or more complexity-adjusted goodness of fit criteria. These criteria, several of which are discussed below, reward goodness of fit, measured in some particular manner, and penalise complexity (or reward simplicity). For any of the following statistics, its value is calculated for each linear model being considered, and then these values are compared; if a model has a particularly good value of this statistic relative to the other models, then this might encourage the analyst to lean towards that model.

### **Adjusted- $R^2$**

We have already mentioned adjusted- $R^2$  several times above: in fact, we previously stated that its formula is

$$\text{adjusted} - R^2 = 1 - \frac{\hat{\sigma}^2}{s^2}$$

and that, the closer it is to 1, the better (in some sense) we consider the model to be. Note that adjusted- $R^2$  can have a negative value, unlike  $R^2$ , which is always between 0 and 1.

We also mentioned that adjusted- $R^2$  is roughly equal to  $R^2$ , which is always larger for more complex models, minus a penalty for more complex models (where complexity is measured by the number of explanatory variables in the model). This fact can be seen by examining an alternative formula for adjusted- $R^2$ :

$$\text{adjusted} - R^2 = \frac{1}{n-p} \{(n-1)R^2 - (p-1)\},$$

where  $p$  is the number of explanatory variables in the model. As can be seen in this formula, for two models with the same  $R^2$  value, the less complex model (with fewer explanatory variables) will have the larger adjusted- $R^2$  value. For two models with the same level of complexity (the same number of explanatory variables), the model with the higher  $R^2$  value will have the larger adjusted- $R^2$  value. The adjusted- $R^2$  statistic tries to achieve a balance between goodness of fit (as measured by high  $R^2$ ) and model simplicity (as measured by small  $p$ ).

### **$C_p$**

Sometimes referred to as *Mallows'  $C_p$* , the  $C_p$  statistic is yet another statistic that

tries to achieve a balance between the often competing goals of goodness of fit and model simplicity. Mallows'  $C_p$  does so by looking at another famous trade-off in statistics: the trade-off between bias and variance/error (for estimators). Essentially, the estimators in more complex linear models are often less biased (which is good), but have more error (which is not good) than the estimators in less complex models. Thus, Mallows'  $C_p$  rewards low bias and also small errors, which is akin to rewarding goodness of fit and simplicity, respectively. In the case of Mallows'  $C_p$ , a smaller value indicates that a model is better (at least in some sense). Typically, a model is looked upon favourably if its  $C_p$  value is close to or below its number of explanatory variables.

### AIC and BIC

When least squares fitting is used for a linear model, it is common to use  $R^2$  to assess goodness of fit. However, sometimes, maximum likelihood fitting is used for a linear model instead of least squares fitting; note that these two types of fitting produce the same results under the assumption that the errors are normally distributed. With maximum likelihood fitting, goodness of fit is often measured using the maximised value of the log-likelihood (known as the *log-likelihood statistic*), for which bigger is better, or the negative of this maximised value (the *negative log-likelihood statistic*), for which smaller is better, instead of  $R^2$ . However, as is the case with  $R^2$ , the log-likelihood statistic always increases (and the negative log-likelihood statistic always decreases) when explanatory variables are added into the model, provided the already included variables do not change; this means that the log-likelihood statistic favours more complex models. For this reason, various complexity-adjusted versions of the log-likelihood statistics have been proposed; these versions differ depending on the specific penalty given to complexity. Two of the most popular complexity-adjusted (negative) log-likelihood statistics are *Akaike's Information Criterion (AIC)*,<sup>1</sup> which adds the amount  $2p$  to the negative log-likelihood as a penalty for model complexity, and Schwarz's *Bayes Information Criterion (BIC)*, which adds the amount  $p \log(n)$  to the negative log-likelihood as a penalty; for both of these statistics, a smaller value indicates a better model (in some sense). Note that the complexity penalty in BIC is larger for bigger data sets, which is not the case for AIC, where the penalty does not change with data set size. The reason that BIC has the penalty increase with sample size is that large data sets are particular culprits in terms of leading one to believe (spuriously) that many explanatory variables affect the response (i.e., are significant) and should be included in the model. Thus, BIC tries to avoid the inclusion of explanatory variables that do not affect the response variable, making it a good criterion to use when a parsimonious model is desired, such as when explanation is the goal. On the other hand, AIC tries to include all variables that do have an effect on the response (possibly including some others that don't in the process), making it a good criterion to use when a well fitting model is desired for prediction.

---

<sup>1</sup> Note that Mallows'  $C_p$  is approximately equal to AIC, giving very similar results, and can be viewed as a sort of complexity-adjusted log-likelihood statistic.

**How to use the above model comparison criteria**

For every linear model being considered, one could use a computer to calculate the value of each of the above criteria. One could then present this information in a table, where rows correspond to linear models and columns correspond to the various criteria. Then, one could scan each column, looking to see which model had the best value for that particular criterion. Hopefully, one model would be rated the best by all criteria, but, if this were not the case, then some decisions would have to be made.

**Example #1 (continued)**

Returning to our Fuel Consumption example, suppose that we are interested in selecting one of four different models (with four different sets of explanatory variables). Note that all categorical variables in these models are treated using the factor approach. For these models, we could make the table suggested above:

Model (Explanatory Variables)	Adjusted-R <sup>2</sup>	C <sub>p</sub>	AIC	BIC
Weight	0.708	1.25	474.6	482.2
Weight, Type	0.738	3.20	469.2	489.5
Weight, Cylinders	<b>0.783</b>	<b>0.543</b>	<b>451.5</b>	<b>471.8</b>
Weight, Horsepower, Drive Type	0.704	3.56	478.7	493.9

In the above table, the best value in each criterion column is bolded and italicised. The linear model with Weight and Cylinders as explanatory variables for Fuel Consumption is City is the best according to the four criteria, which might lead us to select this model, as long as it also appears that the linear model assumptions are satisfied for this model ■

Alternatively, if the analyst wants to use the C<sub>p</sub> criterion to pick a model from a sequence of nested models (each one adding one explanatory variable to the previous model), the information is often presented in a graphical form in a *C<sub>p</sub> plot*. Essentially, the C<sub>p</sub> values for the models are plotted against their number of explanatory variables, and the line at C<sub>p</sub>=p is often drawn in. Then, the model with the smallest C<sub>p</sub> value, hopefully lying below the line, is selected. (See Ramsey and Schafer (2001), *The Statistical Sleuth*, p. 357, for an example of a C<sub>p</sub> plot.)

**Arriving at a model (or at a set of explanatory variables)**

In the previous section, we discussed strategies that can be used to choose between a reasonable number of models. These models have been selected by the analyst as potential candidates for the final model presumably because they appear to

satisfy the linear model assumptions and also to fit the data well while being reasonably simple. One might select from among these models using any of the various comparison statistics discussed above: one might compare them using adjusted- $R^2$ ,  $C_p$ , BIC, or AIC or, if they form a sequence of nested models, F-tests.

However, if there is a large number of potential candidate variables, then the number of models that the analyst must sort through to arrive at a small number of candidate models is enormous. In this situation, the analyst may want to employ an *automated search procedure* with one of the aforementioned comparison statistics to find his final model. Essentially, these procedures use the selected comparison statistic to pick one of the models that falls between a *minimum model* (typically the model with just an intercept and no explanatory variables) and a *maximum model* (typically the model that includes all potential candidate explanatory variables).

There are several different automated search procedures. *Forward selection* begins from the minimum model and adds variables one at a time; at each step, the procedure adds the variable for which the selected comparison statistic ( $F$ , AIC, BIC,  $C_p$ ) is best, and the procedure stops adding variables once doing so does not significantly improve the statistic. The model at which the procedure stops is the final model. Not surprisingly, *backward selection* uses a similar approach, starting with the maximum model and removing variables until it arrives at a final model, and *stepwise selection* starts with some model (typically somewhere between the minimum and maximum models) and alternates between adding and removing variables until a final model is reached. However, the *best subsets* procedure does not move iteratively through a sequence of models, looking only at some of the models between the minimum and maximum model, as do the previous three procedures. Instead, the best subsets procedure looks at all of the models between the minimum and maximum model (typically ignoring any models that aren't hierarchical) and calculates the selected comparison statistic (e.g., adjusted- $R^2$ ,  $C_p$ , AIC, BIC) for each one; the final model is the one with the best value of that statistic.

Be warned that it is best to avoid using these automated procedures for a variety of reasons, one being that they are sort of a glorified form of data dredging. However, if you are going to use one, do not use forward selection! Instead, stepwise selection with the AIC statistic or best subsets selection with the  $C_p$  statistic are reasonable procedures.

### *A Final Note on Model Selection*

Just as the first thing we should look at when checking our model is whether the estimated values of the coefficients are reasonable, we should always make certain that our final model is sensible given existing theory and our previous experience: does it make sense that the explanatory variables included in our final model would actually affect the response variable?

## Review of the Linear Modelling Process

Let us now try to tie together some of the ideas that we have met so far in order to form a sort of strategy for performing linear regression analysis on a particular data set. When doing so, we must remember that the goal of statistical modelling is typically to produce a simple, interpretable model that seems plausible both in light of the data and previous knowledge. The following points outline a possible linear regression analysis strategy:

1. Once we have our data set in a statistical software package, it is a good idea to produce a scatter plot of the response variable versus each potential candidate explanatory variable. This enables us to get some sort of feel for what is happening in the data, albeit only one explanatory variable at a time.
2. We now start fitting models. It is often best to start with a simple model that may include just one or two explanatory variables.
3. Assess the quality of this simple model: test the regression coefficients to see if they are different from 0; check to see whether the assumptions are satisfied; check the model's goodness of fit; and check for outliers.
4. If you don't have a good model (e.g., the assumptions are not satisfied, the model does not fit well, there are worrying outliers), start to look for more complicated models. Fit models that includes new explanatory variables and/or variables derived from those already in the model. Again, check the quality of each model by performing the checks listed in (3.).
5. Hopefully, you have found some models for which you are satisfied with the results of the quality checks. If this is the case, use the model selection techniques described in the "Model Selection" section to pick a final model.
6. If you are not happy with the results of the quality checks for any of the models examined, you might want to consider, for the most reasonable of the models examined, using an alternative fitting procedure (e.g., Generalized Least Squares, robust estimation, resistant estimation) or transforming the response variable. Alternatively, if the situation looks particularly dire, you might wander to abandon the use of linear models altogether and use another sort of model.
7. If you have arrived at a final linear model, see if the results are consistent with your previous knowledge of the phenomenon being modelled: does the model you have fitted make sense, are the regression coefficients the right sign, does it agree with what has already been published in the literature?

### A final example: diamond ring pricing

This example concerns the relationship between the price and various aspects of diamond quality for ladies' diamond rings.<sup>2</sup> Our aim is to discover a plausible relationship that can be used to price the diamond rings. Here, we will consider only 20 carat gold (20K) rings. Thus, the price of the rings will not depend on the gold quality and will only be affected by the stone's four Cs: the carats, cut, colour and clarity of the diamond stone. However, in this example, we will use only Diamond Carats to predict Price: we will consider only Diamond Carat, and various derivatives of it, as explanatory variables. There are 48 20K rings in our data set: the weights of their diamonds range between 0.12 and 0.35 carats (1 carat = 0.2g), and their prices range between \$223 and \$1086.

Let us consider fitting a linear model to Price and Diamond Carats. As per point (1.) in our suggested strategy, let us first look at a scatterplot of the two variables in this data set:

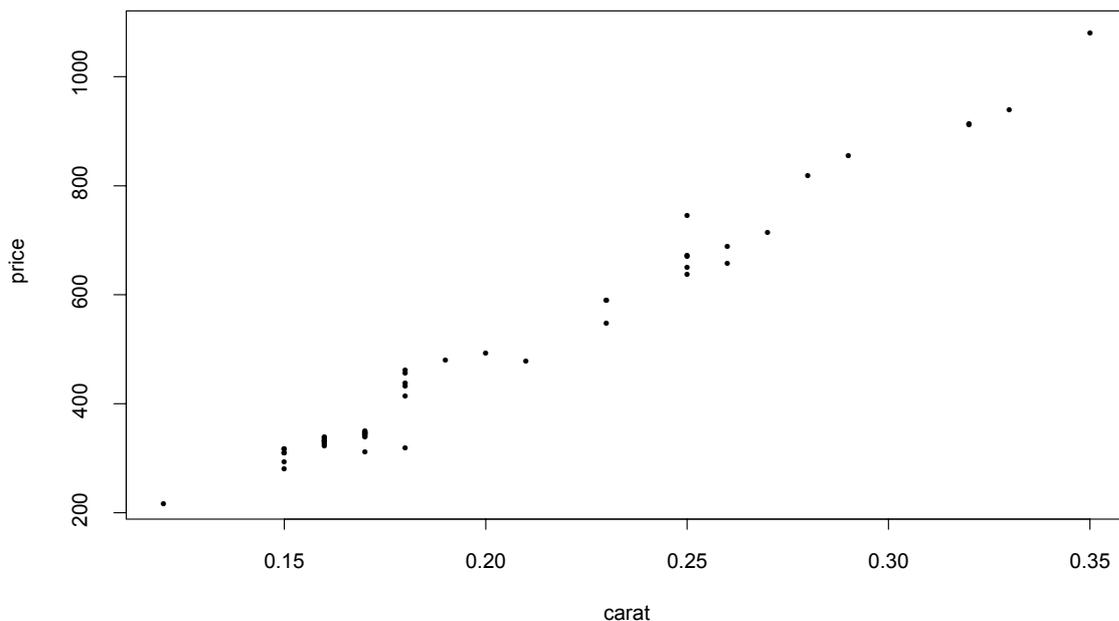


Figure 2: Price versus Diamond Carats for forty-eight 20K rings

It would appear that there is a nice linear relationship between the carat of a diamond ring and its price; therefore, it is reasonable to consider fitting a linear model relating the two. Doing so gives the following results:

Coefficients:

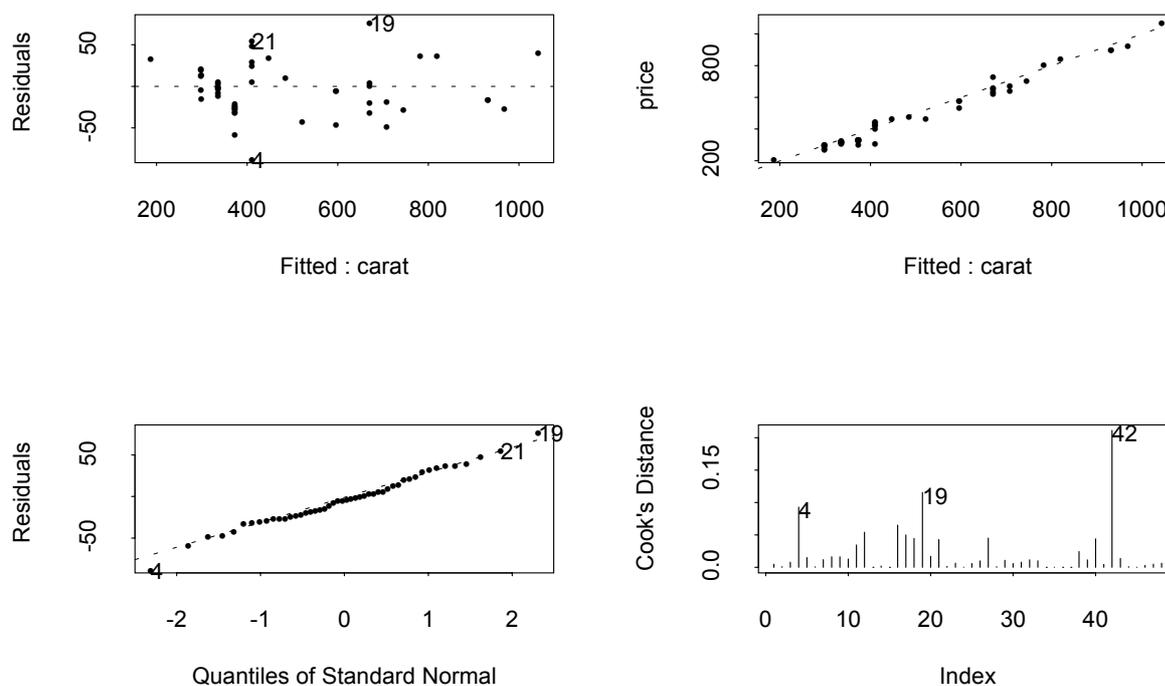
	Value	Std. Error	t value	p-value
--	-------	------------	---------	---------

<sup>2</sup> The data comes from a full-page advertisement placed in the *Straits Times* newspaper on Feb 29, 1992, by a Singapore-based retailer of diamond jewellery. This analysis follows that of Singfat Chu of National University of Singapore as published in *Journal of Statistics Education* v.4, n.3 (1996).

(Intercept)	-259.6259	17.3189	-14.9909	0.0000
Diamond Carat	3721.0249	81.7859	45.4972	0.0000

The  $p$ -value for Diamond Carat is very small, suggesting that its coefficient is significantly different from 0.

Next, we consider the usual diagnostic plots:



**Figure 3:** Diagnostic plots for the linear regression of Price on Diamond Carat

These four plots do not appear to have anything in particular wrong with them: in the upper two plots, there are no apparent patterns in the residuals, the normal distribution appears to be a reasonable assumption based on the lower left plot, and the Cook's distances in the last plot are very small and do not indicate any problems with outliers and influential points.

Our final quality check involves looking at the  $R^2$  statistic, which, for this model is 0.978: 98% of the variation in Price has been explained by Diamond Carat. (Since we have a very simple linear model, we won't worry about using a complexity-adjusted measure of goodness of fit.)

Are we satisfied with the model that we have just fitted? Based on the quality checks just performed, we should be. However, as per point (7.) in our strategy list, we should only accept a model as final if it makes sense based on our previous knowledge. Does this model make sense? No! Looking at the estimated linear model equation,

$$\text{Price} = -259.63 + 3721.02 \times \text{Carat},$$

we see that the estimated  $\beta_0$  (intercept) parameter has a value of -259.63. This is not sensible because it suggests that a zero-carat diamond ring has negative economic value, which cannot be true in general and is particularly untrue because of the way in which the previous model prices the rings. Because of the way in which we are considering pricing these rings, the intercept represents the value of the gold ring (without the stone), which consists of the value of the gold content plus a craftsmanship fee. Thus, it should be non-negative. We can now take one of two approaches:

1. Do nothing on the grounds that we have not observed any Prices close to 0 and, thus, should not try to predict what would happen around 0 since it is therefore extrapolation.
2. Attempt to transform Price and/or Diamond Carat. One possibility is to use the log transformation of the response, which would guarantee a non-negative predicted Price for a Diamond Carat value of zero.

We consider fitting the same model as before but using  $\log(\text{Price})$  rather than Price as the response variable. Doing so gives us the following output and diagnostic plots:

Coefficients:

	Value	Std. Error	t value	p-value
(Intercept)	4.7489	0.0477	99.4640	0.0000
Diamond Carat	6.7872	0.2255	30.1027	0.0000

R-Squared: 0.9517

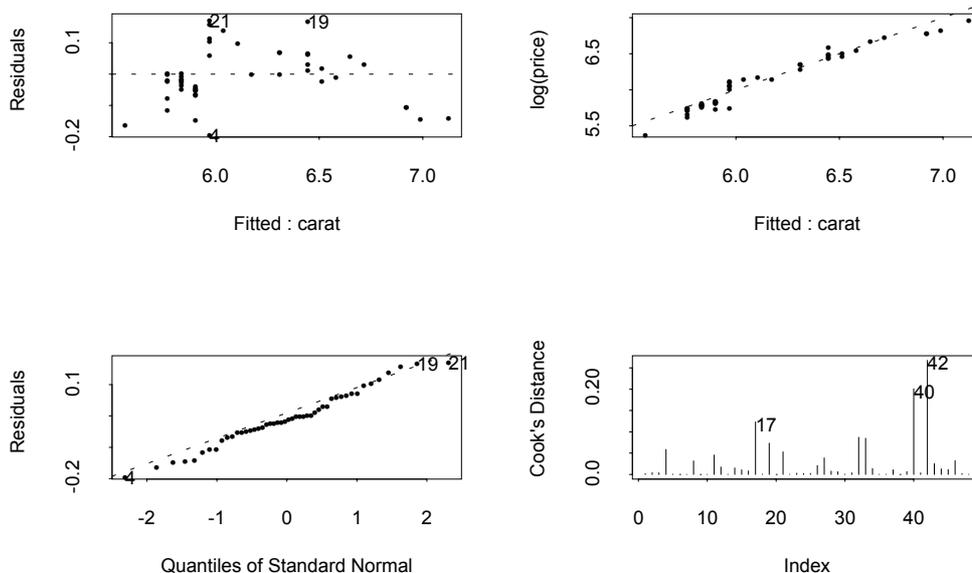


Figure 4: Diagnostic plots from the regression of  $\log(\text{Price})$  on Diamond Carat

Although the model is a good one in terms of the reasonability of its coefficients

and their high degree of significance, there is a problem with the diagnostic plots. In the upper left plot, the residuals are scattered around a straight line but with a pattern: the residuals follow an upside-down U-shaped pattern as Diamond Carat increases. This suggests that our most recent model might be missing some degree of curvature with respect to carat and that we should consider a model that incorporates the curvature. The obvious choice is a model that includes a (Diamond Carat)<sup>2</sup> term as well as (Diamond Carat). If we finally fit this model to the data we obtain:

Coefficients:

	Value	Std. Error	t value	p-value
(Intercept)	3.8872	0.1605	24.2250	0.0000
Diamond Carat	14.8597	1.4726	10.0909	0.0000
(Diamond Carat) <sup>2</sup>	-17.5370	3.1762	-5.5214	0.0000

R-Squared: 0.9712

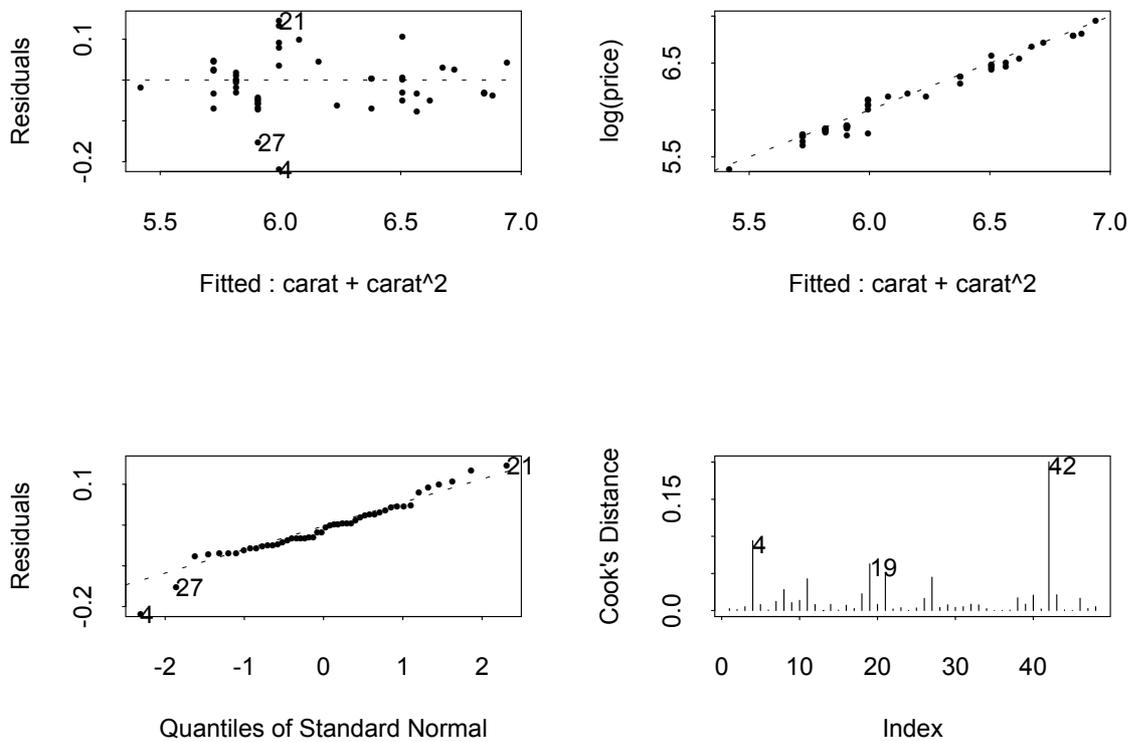


Figure 5: Diagnostic plots from the regression of Log(Price) on Diamond Carat and (Diamond Carat)<sup>2</sup>

This model appears to fit very well, as indicated by its high  $R^2$  value of 0.97, and to satisfy the assumptions, as indicated by the diagnostic plots. Further, its estimated coefficients are sensible. The estimated equation for the final linear model is:

$$\text{Log}(\text{price}) = 3.89 + 14.86 \times \text{Carat} - 17.54 \times \text{Carat}^2.$$

---

*A. Roddam (2000), K. Javaras (2002), and W. Vos (2002)*