1.5 Model Evaluation I (Fit of Functions)

We have used our knowledge of problem solving and properties of linear functions to *develop* models to help communicate solutions to several problems, like identifying the height of an enemy operative, determining the size of cadets for uniform issue, and calculating the equilibrium price of a supply and demand system. Developing models is useful, but our models are limited if we do not know how good they are. This section will focus, not on the development of models, but on their *evaluation.* We will discuss how to determine a model's "goodness," thereby putting more credibility behind the predictions we make.

1.5.1 Model Evaluation (Subjective)

So far, we have informally used two tests to determine our model's goodness: the *nature of the data* test and the "eyeball" test. We will now formally define these tests.

- Nature of the Data: The underlying structure of the data; i.e., how the data would appear if collected in a "perfect world." For example, we would expect projectile motion to follow a parabolic trend as gravity acts on the projectile.
- Eyeball Test: A qualitative measure that determines how closely a model appears to fit a given data set.

We may consider a linear model to be "good" if we knew about the circumstances under which the data were collected and expected the modeled quantities to be proportional – a measure that the model matches the nature of the data. If we graphed the model and the prediction looked like it closely represented the data (passes the eyeball test), then we may call it a "good" model, like the models presented in Figure 1.20. Today, we will develop a means to quantitatively evaluate just how good a model is. How can we tell if one model is really better than another? We will explore this question and more in this lesson.

Let's consider the average demand, the model we discussed in Section 1.1, and the two demand models we developed in Section 1.4.

$$\begin{aligned} &Average_Demand: && y = 15.21 \\ &Estimated\ Linear\ Model\,1: && y = -0.1263x + 26.695 \\ &Estimated\ Linear\ Model\,2: && y = -0.148x + 28.52 \end{aligned}$$

It appears in Figure 1.20 that the average does not "fit" the data well. In fact, the model underpredicts for the first half of the data and overpredicts for the second half of the data. Both estimated linear models seem to be a reasonable fit of our data set. Which line is the better linear model most accurately predicting oil demand given the price of oil per barrel? Both the solid and dashed lines seem to provide a good approximation, but they clearly are different lines (models). So which model best fits the data? We need to define a measure for best fit so that we can compare the models we developed.

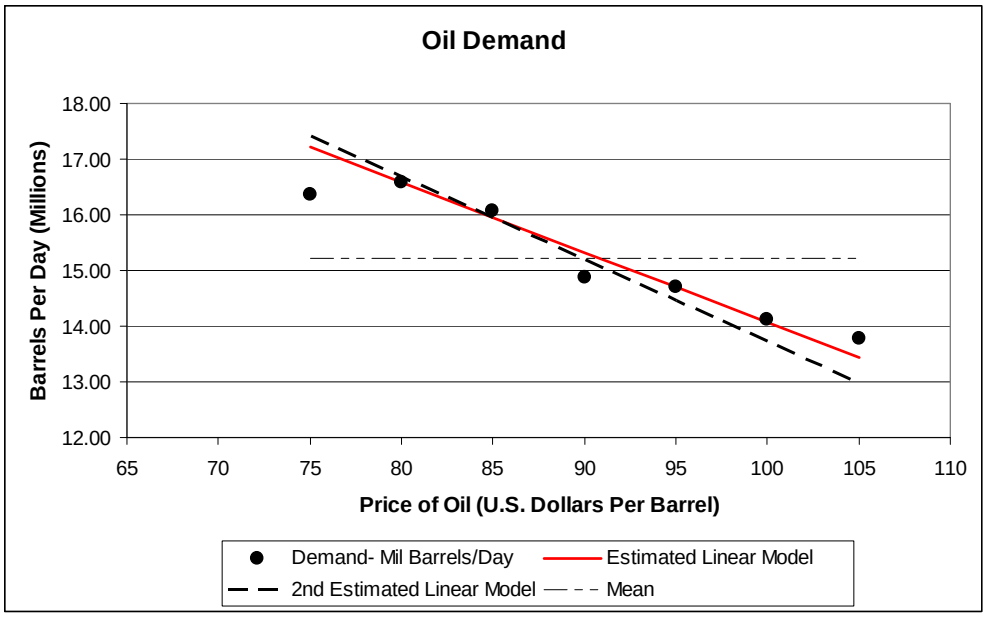


Figure 1.20: A Comparison of Linear Oil Demand Models

1.5.2 Model Evaluation (Sum of Squared Error)

One way to determine which model is "*best*" quantitatively is to measure the distance between the values predicted by the model, $\hat{y}$, and the actual data points, y. This distance is our error, e. (This symbol, *e, is not to be confused with the number, e ≈ 2.72).*

Let's look at the situation a bit closer. If we look at a single data point from Figure 1.20, say, (95 dollars, 14.72 million barrels) we see that both estimated linear models under-predict the actual amount of oil in demand. This is because both models lie slightly below the point. Does this mean that our models are wrong? Absolutely not! The models still capture the general trend of the data; they just predict that consumers will demand less oil, given the oil's selling price.

So, how good or bad are our models? To *quantify,* or put a numerical value on how good or bad they are, we measure the distance between our predictions and the actual data points (see Figure 1.21). In this case, our demand for oil is 14.72 million barrels our solid model predicts a demand of about 14.70 million barrels.

What is the error of the solid model? Well, it's $\textit{actual demand} - \textit{predicted demand}$, or in this case, it is $14.72 - 14.70 \approx 0.02$ million barrels.

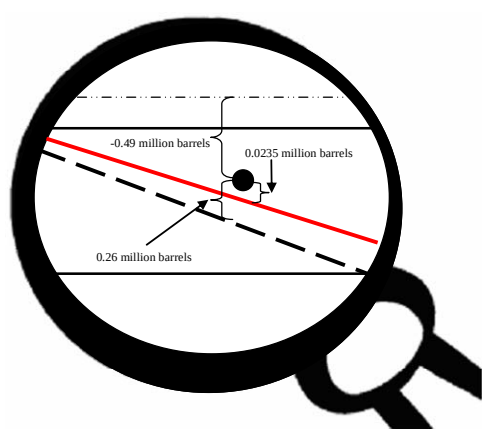


Figure 1.21: Closer Look at Our Errors

This calculation shows that the model under-predicts the actual demand by 0.02 million barrels of oil. What would have happened if we looked at the point just to the left, (90, 14.88). In this case, our model over-predicts; the error is $14.88 - 15.328 = -0.448$ million barrels.

So, what do we do with this information? How can we come up with a single number that *quantifies* how good or bad the entire model is? Let's look at a couple of possibilities.

<u>A-Not-So-Good-Idea:</u> If we sum the errors for each model, perhaps the model with the lowest sum would be the better of the two models. Notationally, with n points in a given model, this is written:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} e_i$$

The above notation may be a little intimidating at first. If we read it from the inside out, it may become a little easier to understand. The quantity $(y_i - \hat{y}_i)$ means "the actual data point minus the predicted value." The subscript i helps us to keep track of which data point we are calculating. For example, $(y_1 - \hat{y}_1)$ is the calculation of the error for the first data point, where $(y_2 - \hat{y}_2)$ is the second error calculation. The symbol $\sum_{i=1}^{n}$ means "the sum from the first value to the nth (n is the final value). So, putting it all together, $\sum_{i=1}^{n}(y_i - \hat{y}_i)$ means "the sum of errors, from 1 to n."

There is one major problem with this technique. If we assume a positive distance when the data point is above the line and a negative distance below the line, a value of zero may only mean that we overestimated and underestimated exactly the same amount. A value of zero would occur because the positive and negative errors would cancel each other out. We really want to measure the total deviation of the model from the data which this method does not capture.

<u>A-Bit-Better-Idea:</u> Perhaps we could take the absolute values of the distances. Notationally, with n points, this is written:

$$\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| = \sum_{i=1}^{n}\left|e_i\right|$$

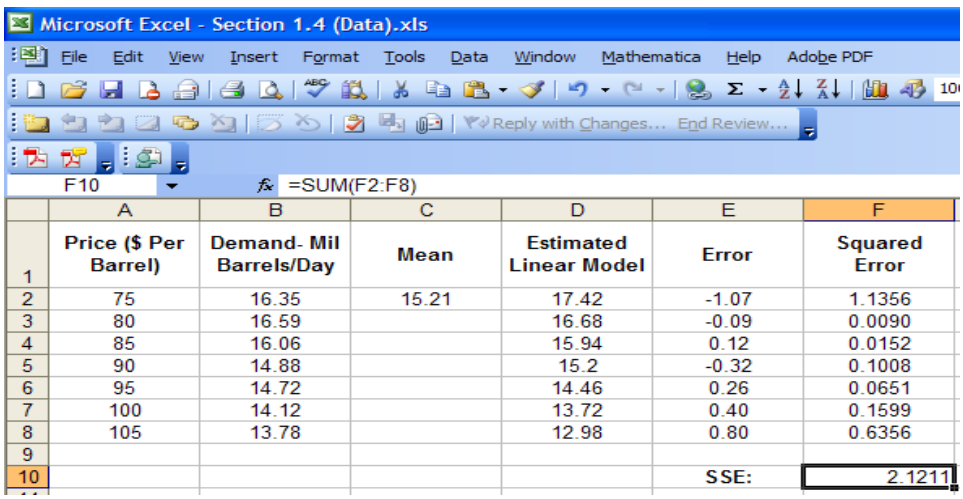
Using absolute values may seem reasonable, but when absolute values are involved, trying to minimize the error between the model and the actual data using calculus techniques becomes quite difficult.

<u>A-More-Acceptable-Idea:</u> A more common and acceptable method is to square the distances before we sum them. Notationally, with n points, this is written:

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

This accomplishes two purposes. First, it eliminates any negative errors (that might cancel out positive errors), and second, it emphasizes those data points that are furthest from our model. That is, we are severely penalized for data that are far from our model when the distance is squared. The *value of summing the square of the errors* is called the **Sum of Squared Errors** (*SSE*). If we compare two models using this method, the one with the smaller *SSE* indicates that the model "fits" the data better than the one with the larger *SSE*. We will consider other model evaluation methods in future lessons.

How can you calculate the *SSE* using MS *Excel*? First, create two columns next to your data and model to track the error and squared error terms. Figure 1.22 demonstrates one way to set up an MS *Excel* spreadsheet.



	A	B	C	D	E	F
1	Price ($ Per Barrel)	Demand- Mil Barrels/Day	Mean	Estimated Linear Model	Error	Squared Error
2	75	16.35	15.21	17.42	-1.07	1.1356
3	80	16.59		16.68	-0.09	0.0090
4	85	16.06		15.94	0.12	0.0152
5	90	14.88		15.2	-0.32	0.1008
6	95	14.72		14.46	0.26	0.0651
7	100	14.12		13.72	0.40	0.1599
8	105	13.78		12.98	0.80	0.6356
9						
10					SSE:	2.1211

Figure 1.22: Sample MS *Excel* Spreadsheet

Given a linear model with a slope of -0.148 and an intercept of 28.52, the model portrayed in Figure 1.22 displays an SSE value of about 2.1211. Is this good? What can we do with this value? SSE is a ***relative*** number. This means that it is only useful when we can compare it with the SSE values from other models.

Think about it this way – suppose that we had a data set that had one million points. By some stroke of genius, let's say you found a model that was only 0.1 units off from each of the one million data points. To the naked eye, the data and your model appear to be exactly the same. The error for each point is only 0.1 and the squared error for each of these points 0.01. Yet the *SSE* for this model is:

$$1{,}000{,}000\ data\ points \cdot (0.1\ units)^2 = 10{,}000\ units^2$$

What does this tell us? Is this a large or small number? Unfortunately, an SSE by itself does not tell us much. Without comparing this SSE to the SSEs of other models, we cannot tell if the SSE above is high or low.

Question 1 *Figure 1.22 shows the calculation of the SSE for the second estimated linear model, with parameters 'a'=-0.148 and 'd'=28.52. Is it better than the first estimated linear model developed in Section 1.4, with parameters 'a'=-.1263 and 'd'=26.695?*

Question 2 U*sing the Space Shuttle Challenger data set from Section 1.4, create a spreadsheet like the one shown in Figure 1.22. What is the SSE for the model estimated in Section 1.4? If you change the values of the slope and intercept for this model, can you get a smaller SSE?*

1.5.3 Model Evaluation (Coefficient of Determination)

Another quantitative measure commonly used in statistics is the coefficient of determination, or r^2. This value is bounded by zero and one ($0 \le r^2 \le 1$), and is ***a measure of the amount of variation of the data for which your model can account.*** So, an $r^2 = 0.83$ would mean that 83% of the variation of the data can be accounted for by your model, the remaining 17% of the variation must be the result of some other factors not covered in your model.

> The **coefficient of determination** or r^2 is a measure of the amount of variation of the data accounted for by the model and is bounded by zero and one ($0 \le r^2 \le 1$).
>
> $$r^2 = 1 - \frac{SSE}{SST} \qquad (1)$$

Before we calculate r^2, we must define a key term, the Sum of Squares Total, SST. Recall that the Sum of Squared Error was a measure of how the model you developed compared to the data for which it was developed. SST is similar, but instead of calculating the error between the data and a model you estimated, SST calculates the error between the data and the mean – the simplest model in our toolkit.

The mean is a simple model with just one parameter, so we expect it to be our least accurate model. Because SST quantifies the error of the least accurate model, we can say that it is the measure of how the worst model, the mean, compares to the data for which it was developed. SST is calculated in a similar manner to SSE,

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

where SST is the sum of squared deviations of the y-values of the data from the sample mean of the data. The sample mean is the average of the given outputs denoted $\bar{y}$. SST is the relative measure of how bad the mean is as a model, that's why it's called Sum of Squares Total – it's the total amount of squared error using the worst model.

If we divide SSE (the error between the data and the model you developed) by the SST we get the percentage of error in the data that our model does not explain.

The best possible model would have zero error ($SSE = 0$), accounting for 100% of the variability in the data. If we subtract the percentage of error that our model does *not* account for from 100% (or 1), we get a measure of error that the model *does* account for. Therefore, $r^2 = 1 - \frac{SSE}{SST}$.

Example 1 *In Section 1.4, you created a linear model for homework that predicted the height of an R-day cadet based on his or her weight. Our best fit model is:*

$$height(weight) = 0.135451(weight) + 47.94905$$
$$\text{Domain}: \{weight \mid 90 < weight < 300\} = (90, 300)$$
$$\text{Range}: \{height \mid 60.13 < height < 88.57\} = (60.13, 88.57)$$

It has an SSE of approximately 1.95. Now, we will turn our analysis to r^2 to determine the accuracy of our model. Recall, we will need to calculate SST using Equation (2) on the previous page where SST is calculated by:

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

Since we determined the SSE of this model previously, we focus here on how to calculate the SST.

- First, we must determine the sample mean which is the average of the data values; here it is the average height. The average height is 70.6 inches.
- Next, we determine how far each of the data values is from this mean - that is, determine the deviations of the y-values of the data from the sample mean of the data.
- Then, we square each of these values.
- Finally, we sum them up.

The SST for this example is 80.4.

The final step in determining the r^2 is to use Equation (1) inputting our SSE and SST where appropriate.

$$r^2 = 1 - \frac{1.94705456}{80.4} = 0.975783$$

Here, our model accounts for 97.6% of the variation in our y-values. This indicates that our model is very accurate for this data set.

One attribute of r^2 is that, unlike SSE, it is a standardized or normalized measure. In other words, r^2 values do not need to be compared to substantiate the use of one model over another. If your model has a high r^2 it indicates it is accurate for the set of data you are modeling.

Caution! A high r^2 value *does indicate the model is accurate* for the data *but does not necessarily indicate that it is appropriate* for the situation being modeled. Similarly, a low r^2 (say in the low 40s) does not necessarily mean the model is not useful. For example, we can find polynomial models to fit every data set but not every data set should be modeled with a polynomial function. To ensure a model is both accurate and appropriate, employ more than one evaluation tool.

1.5.4 Model Evaluation (Number of Parameters)

An important consideration when modeling is the relationship between the number of parameters in a model and the measure of 'goodness' (often SSE or r^2). Every time we add a new parameter to a model, the SSE will get smaller and r^2 will always get larger, even though the added parameter may end up being of insignificant value. The question for modelers is, "when are enough parameters enough?" It's often best to use this notion of Albert Einstein, "Keep things as simple as possible, but no simpler."

Let's explore an example that will show use the effect of adding additional terms and parameters to a model.

Example 2 *In the following graph, Figure 1.23, we once again see the seven points of data that define the oil demand problem, this time with three models overlaid on the data.*

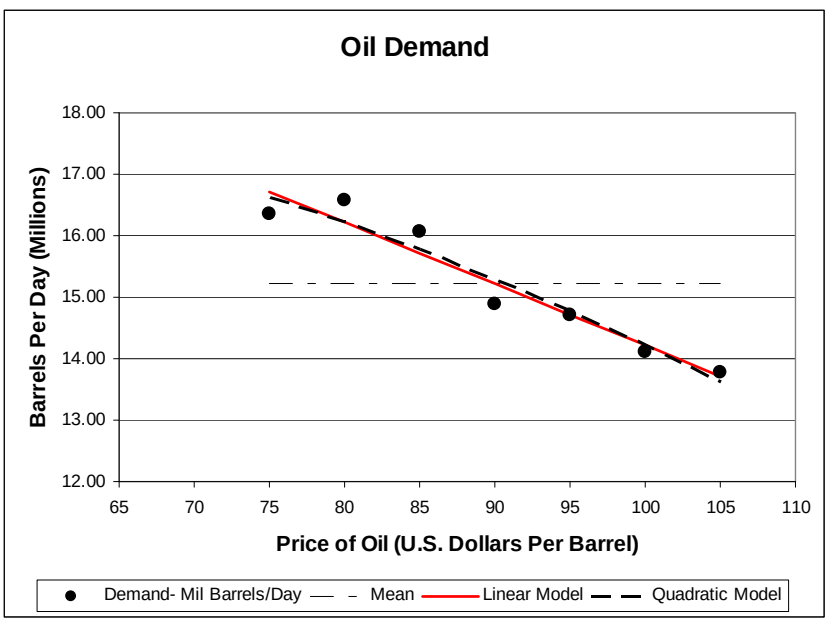


Figure 1.23: Oil Demand Modeled with Three Models

Clearly, the relationship between the independent variable and the dependent variable is decreasing. Let's model the data with polynomials having one, two, and three parameters (mean, linear with non zero y-intercept, and quadratic). NOTE: To estimate the 3-parameter model, we chose three representative data points and used a polynomial of the form $f(x) = a(x+c)^2 + d$ *, which you'll see again in section 1.8. Table 1.7 lists the model, minimum SSE, and corresponding parameter values for the data.*

parameters	*model*	*Min SSE*	*a*	*c*	*d*
1	*d*	*7.52*			*15.21*
2	$f(x) = ax + d$	*0.49*	*-0.1263*		*26.695*
3	$f(x) = a(x+c)^2 + d$	*0.51*	*-.0053*	*-81.99*	*16.611*

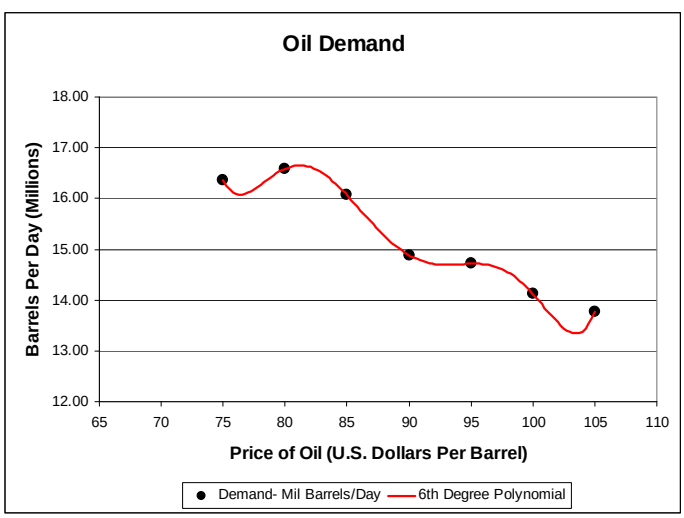
Table 1.7: Polynomial Models

From the graph and the table, we can note a few important things. First, the minimum SSE for the one-parameter model is an order of magnitude higher – has a change of more than 10 times – than the two-parameter model (a linear model). But, the improvement of going to a three-parameter model (a quadratic model) is only about 4%. Also, graphically the one-parameter model is clearly poor, while the other two are virtually indistinguishable from each other. Finally, then the third parameter, a, is introduced, its value (a = -0.0053) is so close to zero when compared to the scale of the other parameters, that it is not a useful addition to the modeling process.

Because the addition of the third parameter is not useful in this case, we would decide to use the simpler linear model. However, we noticed that the SSE decreased when we added a parameter to model, a fact that will always hold true. A larger number of terms will *always* result in lower SSE (and larger r^2). **In fact, a polynomial of degree $n - 1$ will always fit n data points.** For example, a data set of two points can be fit perfectly by a degree 1 polynomial (a line). Also, three points can be fit by a 2nd degree polynomial (a quadratic). Each would have a SSE = 0. A polynomial of degree n - 1 will always fit n data points. But if a polynomial of degree $n - 2$ can closely fit the data points, why use a more complicated model?

In Example 2, there were 7 data points. A 6-degree polynomial can model these points "perfectly," with an SSE of zero and r^2 of 1. A 6-degree polynomial is a very "wiggly" function (see picture at right). As you would expect, the function is also very complicated:

$$f(x) = .0000009x^6 - .0005x^5 + 0.1066x^4$$
$$-12.692x^3 + 847.71x^2 - 30102x + 444020$$



Notice that the model's trend is opposite of the data, seeming to indicate that for prices higher than $105, demand will increase. It doesn't make sense that higher price would yield higher demand. This model, though it fits every data point, is not going to be useful for prediction.

As mentioned previously, in every case, as you add a parameter, SSE will decrease. But the important question is, "does the improvement in SSE justify the disadvantage of adding a parameter and increasing the complexity of the model?" The nature of the disadvantage of having too many parameters lies in the strength of the estimates for those parameters. Every parameter that needs to estimated essentially 'takes away' some of the information in the sample. Estimating two parameters from seven data points will be fairly easy and will result in relatively accurate estimates for those two parameters. On the other hand, estimating 6 parameters from 7 data points is far more complicated and will likely result in parameters of little consequence, such as *a* from Table 1.7. Furthermore, once you have more parameters than data points, it is not possible to estimate any of the parameters.

What's a good rule of thumb in determining how many parameters to add to your model? When adding an extra parameter doesn't give noticeable graphical improvement or a noticeable improvement on SSE, then stop adding parameters. What is noticeable? This is a tough question and often depends on the situation…another example of why modeling has a "science" component and an "art" component. Remember: for the model you develop to have meaning, you must be able to interpret its results to a decision maker. If you are unable to communicate the effect of an additional parameter, it's a good bet that you may not need that parameter!

Question 3 *You are working with a group of student interns for the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, monitoring the global warming situation. NCAR has access to data about* CO_2 *concentrations in the atmosphere (below and on course website). Let x represent the number of years since 1995.*

years	0	1	2	3	4	5	6	7	8	9
CO_2 ppm	361.6	364.0	364.6	367.3	369.6	370.5	372.1	373.5	376.1	379.0

a. *Develop a linear model that fits these data. Show all work in how you obtained your model and put a graph of your model on the same axes as the scatterplot.*

b. *Select two different points to estimate a linear model.*

c. *Thoroughly evaluate your model's, meaning: use the nature of the data test, the eyeball test, SSE,* r^2*, and number of parameters. Is your model a good fit for the data?*

d. *These data were all observed in March of each year. Based on your model what were the* CO_2 *ppms in September 2001?*

e. *Based on your model what is your prediction for the* CO_2 *ppm in March 2100?*

f. *The equation used at NCAR to predict CO_2 ppm was* $y_2 = 1.82x + 361.6$. *Plot this equation with your model and the data on your scatter plot. Which equation seems to be the "better" model? Explain your reasoning.*

Question 4 *Using the models developed in Example 1 of Section 1.4 (O-Ring Data), determine the* r^2 *value for each model. Which appears to be the better model? Based on your understanding of the problem, interpret the meaning of your* r^2 *values and explain what it indicates about the best model. Discuss your findings.*

***Question 5** *You are working with a group of student interns for the park services at Yellowstone National Park in Wyoming. Old Faithful is one of 400 geysers within the park. It is named Old Faithful, because the time of next eruption can be predicted by the duration of Old Faithful's eruption. Since the original model was established, the intervals between eruptions have tended to increase. The park service would like you to set up a model with current data and answer a few of their questions with the model. Below are some current data on Old Faithful's eruptions.*[1] *The data is also linked on the course website.*

x duration in minutes	1.8	1.98	2.37	3.78	4.3	4.53	1.82	2.03	2.82	3.83	4.30	4.55	1.88	2.05
y interval in minutes	56	59	61	79	84	89	58	60	73	85	89	86	60	57

[1] *These data were obtained from Yellowstone National Park.*

a. *Based on the graph you develop, does it appear that interval is approximately a linear function of duration?*

b. *What is the slope of the line that models this data? Explain in practical terms (duration and interval) the meaning of this slope?*

c. *What is the y-intercept for your model? Does this intercept have any practical meaning? If so, what is it?*

d. *Thoroughly evaluate the model you developed, meaning: apply the nature of the data test, the eyeball test, and* r^2.

e. *Suppose that you observe an eruption that lasts 2 minutes and 40 seconds. Based on your model, predict when to expect the next eruption.*

f. *The equation used by the park service to predict the intervals between Old Faithful's eruptions is* $y = 14x + 30$. *Compare this equation with your model. Which equation seems to be the "better" model? Explain your reasoning.*