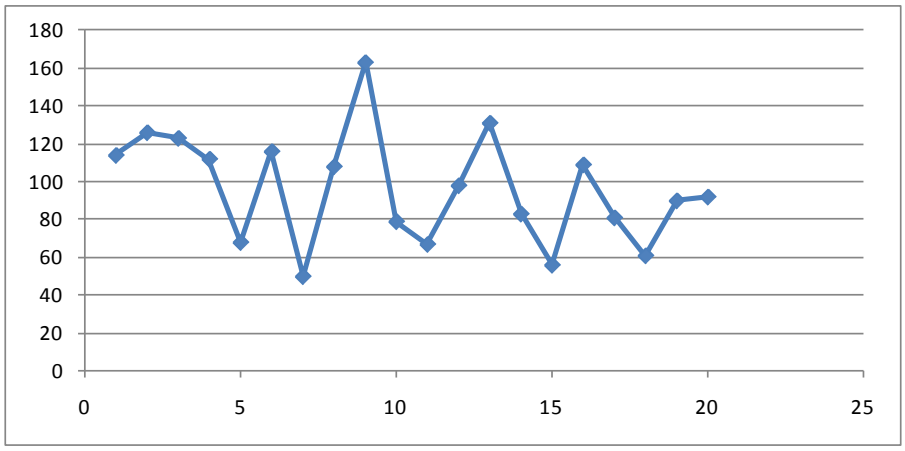
Review of basic statistics and the simplest forecasting model: the sample mean

Robert Nau
Fuqua School of Business, Duke University
August 2014

Most of what you need to remember about basic statistics

Consider a random variable called X that is a *time series* (a set of observations ordered in time) consisting of the following 20 observations:

114, 126, 123, 112, 68, 116, 50, 108, 163, 79, 67, 98, 131, 83, 56, 109, 81, 61, 90, 92.



How should we forecast what will happen next? The simplest forecasting model that we might consider is the *mean model*,[1] which assumes that the time series consists of independently and identically distributed ("i.i.d.") values, as if each observation is randomly drawn from the same population. Under this assumption, the next value should be predicted to be equal to the historical sample mean if the goal is to minimize mean squared error. This might sound trivial, but it isn't. *If you understand the details of how this works, you are halfway to understanding linear regression.* (No kidding: see section 3 of the regression notes handout.)

To set the stage for using the mean model for forecasting, let's review some of the most basic concepts of statistics. Let:

X = a random variable, with its individual values denoted by x_1, x_2, etc.
N = size of the entire population of values of X (possibly infinite)[2]
n = size of a finite sample of X

[1] This might also be called a "constant model" or an "intercept-only regression."
[2] The term "population" does not refer to the number of *distinct* values of X. The same value could occur many times in the population. For example, the values of X could be integers or just 0's and 1's.
 Main web site: people.duke.edu/~rnau/forecasting.htm

The *population* ("true") *mean* μ is the average of the all values in the population:

$$\mu = \sum_{i=1}^{N} x_i / N$$

The *population variance* σ^2 is the average squared deviation from the true mean:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

The *population standard deviation* σ is the square root of the population variance, i.e., the "root mean squared" deviation from the true mean.

In forecasting applications, we never observe the whole population. The problem is to forecast from a finite sample. Hence statistics such as means and standard deviations must be estimated with error.

The *sample mean* is the average of the all values in the sample:

$$\bar{X} = \sum_{i=1}^{n} x_i / n$$

This is the "point forecast" of the mean model for all future values of the same variable. The sample mean of the series X that was shown above is 96.35. So, *under the assumptions of the mean model, the point forecast for X for all future time periods should be 96.35.*

The *sample variance* s^2 is the average squared deviation from the sample mean, except with a factor of $n-1$ rather than n in the denominator:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n-1}$$

The *sample standard deviation* is the square root of the sample variance, denoted by *s*. The sample standard deviation of the series X is equal to 28.96.

Why the factor of $n-1$ in the denominator of the sample variance formula, rather than n? This corrects for the fact that the mean has been estimated from the same sample, which "fudges" it in a direction that makes the mean squared deviation around it less than it ought to be. Technically we say that a "degree of freedom for error" has been used up by calculating the sample mean from the same data. The correct adjustment to get an "unbiased" estimate of the true variance is to *divide the sum of squared deviations by the number of degrees of freedom, not the number of data points*

The corresponding statistical functions in Excel[3] are:

- **Population mean = AVERAGE (x_1, …x_N)**
- **Population variance = VAR.P (x_1, …x_N)**
- **Population std. dev. = STDEV.P (x_1, …x_N)**
- **Sample mean = AVERAGE (x_1, …x_n)**
- **Sample variance = VAR.S (x_1, …x_n)**
- **Sample std. dev. = STDEV.S (x_1, …x_n)**

Now, why all this obsession with *squared* error? It is traditional in the field of statistics to measure variability in terms of average *squared* deviations instead of average *absolute*[4] deviations around a central value, because squared error has a lot of nice properties:

- The central value around which the sum of *squared* deviations are minimized is, in fact, the *sample mean.* This may not be intuitively obvious, but it is easily proved by calculus. So, *when we fit forecasting models by minimizing their sums of squared errors, we are implicitly calculating means, even when we are estimating many things at once.* In particular, when we estimate the coefficients in a linear regression model by minimizing squared error, which our regression software does for us automatically, we are implicitly calculating the "mean effect" of each of the independent variables on the dependent variable, in the presence of the others.
- Variances (rather than standard deviations or mean absolute deviations) are *additive* when random variables that are *statistically independent* are added together.
- From a decision-theoretic viewpoint, *large errors often have disproportionately worse consequences than small errors*, hence squared error is more representative of the economic consequences of error. Why? A small error in your analysis will probably not result in a bad decision or a wrong conclusion. The future may turn out slightly different from what you expected, but you probably would have done the same thing anyway, so it doesn't matter very much. However, if the error is large enough, then it may lead to a wrong-headed decision or conclusion that will have bad consequences for you or somebody else. So, in many situations we are relatively more concerned about the occasional large error than the more frequent small error. Minimizing squared error when choosing among forecasting models is a rough guideline for doing this.
- Variances and covariances also play a key role in *normal distribution theory and regression analysis*, as we will see. All of the calculations that need to be done to fit a regression model to a sample of data can be done based only on knowledge of the sample means and sample variances and covariances of the variables.

[3] In earlier versions of Excel the sample standard deviation and variance functions were STDEV and VAR, and the corresponding functions for population statistics were STDEVP and VARP.

[4] Actually, there is a lot of interest nowadays in the use of absolute error rather squared error as the objective to be minimized when fitting models, especially in econometrics. This approach yields parameter estimates that are less sensitive to the presence of a few large errors and is also useful for model selection in high-dimensional data sets, but it is beyond the scope of this course.

The *standard error of the mean* is:

$$SE_{mean} = s/\sqrt{n}$$

- This is the estimated standard deviation of the *error* that we would make in using the sample mean $\bar{X}$ as an estimate of the true mean μ, if we repeated this exercise with other independent samples of size *n*.
- It measures the *precision* of our estimate of the (unknown) true mean from a limited sample of data.
- As *n* gets larger, SE_{mean} gets smaller and *the distribution of the error in estimating the mean approaches a normal distribution*. This is one of the most fundamental and important concepts in statistics, known as the "Central Limit Theorem."
- In particular, it decreases in inverse proportion to the square root of the sample size, so for example, *4 times as much data reduces the standard error of the mean by 50%.*

What's the difference between a standard *deviation* and a standard *error*?

- The term "standard deviation" refers to the *actual* root-mean-squared deviation of a population or a sample of data around its *mean*.
- The term "standard error" refers to the *estimated* root-mean-squared deviation of the *error* in a parameter estimate or a forecast under repeated sampling.
- Thus, a standard error is the "standard deviation of the error" in estimating or forecasting something

The mean is not the *only* statistic for measuring a "typical" or "representative" value drawn from a given population. For example, the *median* (50^{th} %-tile) is another summary statistic that describes a representative member of a population. If the distribution is symmetric (as in the case of a normal distribution), then the sample mean and sample median will be approximately the same, but if the distribution is highly "skewed", with more extreme values on one side than the other, then they may differ significantly. For example, the distribution of household income in the U.S. is highly skewed. The median US household income in 2010 was $49,445, whereas the mean household income was $67,530, about 37% higher, reflecting the effect of a small number of households with extremely high incomes. Which number better measures the income of the "average" household?

That being said, the most commonly used forecasting models, such as regression models, focus on means (together with standard deviations and correlations) as the key descriptive statistics, and point forecasts are usually expressed in terms of mean values rather median values, because this is the way to minimize mean squared error. Also, in many applications (such as sales forecasting), the total over many periods is what is ultimately of interest, and predictions of mean values in different periods (and/or different locations) can be added together to predict totals.

Another issue is that when forecasting at a very fine level of detail (e.g., units of a given product sold at a given store on a given day), the median value of the variable in a single period could be

zero! Expressing a forecast for such a variable in terms of the median of its distribution would be trivial and uninformative.

Furthermore, nonlinear transformations of the data (e.g., log or power transformations) can often be used to turn skewed distributions into symmetric (ideally normal) ones, allowing such data to be well fitted by models that focus on mean values.

Forecasting with the mean model.

Now let's go forecasting with the mean model:

- Let $\hat{x}_{n+1}$ denote a *forecast* of x_{n+1} based on data observed up to period n
- If x_{n+1} is assumed to be independently drawn from the same population as the sample x_1, …, x_n, then *the forecast that minimizes mean squared error is simply the sample mean:*

$$\hat{x}_{n+1} = \bar{X}$$

In the special case of the mean model, the sample standard deviation (s) is what is called the *standard error of the model*, i.e., the estimated standard deviation of the intrinsic risk. Now, what is the standard deviation of the *error* we can expect to make in using $\hat{x}_{n+1}$ as a forecast for x_{n+1}? This is called the *standard error of the forecast* ("SE_{fcst}"), and it depends on both the standard error of the model and the standard error of the mean. Specifically, it is the square root of the sum of the squares of those two numbers:

$$SE_{fcst} = \sqrt{s^2 + SE_{mean}^2} = s\sqrt{1+\frac{1}{n}} \approx s(1+\frac{1}{2n})$$

SE_{fcst} measures the *forecasting* risk, *assuming the model is correct*

End result: for the mean model, SE_{fcst} is slightly larger than the sample standard deviation

The *standard error of the model* measures the *intrinsic* risk (estimated "noise" in the data); for the mean model, the standard error of the model is just the sample standard deviation

The standard error of the mean measures the *parameter* risk (error in estimating the "signal" in the data)

Note that if you square both sides, what you have is that the estimated *variance* of the forecast error is the sum of the estimated variance of the noise and the estimated variance of the error in estimating the mean.

Variances of the different components of forecast error are *always* additive in this way, for linear[5] forecasting models with normally distributed errors. In fact, we can call this the "fundamental law of forecasting risk:"

Variance of forecasting risk = variance of intrinsic risk + variance of parameter risk

It does not take into account the model risk, though!

For the mean model, the result is that the *forecast standard error is slightly larger than the sample standard deviation*, namely by a factor of about $1+(1/(2n))$. Even for a sample size as small as n=20 there is not much difference: 1+(1/40) = 1.025, so the increase in forecast standard error due to parameter risk (i.e., the need to estimate an unknown mean) is only 2.5%. In general, the estimated parameter risk is a relatively small component of the forecast standard error *if* (i) the number of data points is large in relation to the number of parameters estimated, *and* (ii) the model is not attempting to extrapolate trends too far into the future or otherwise make predictions for what will happen far away from the "center of mass" of the data that was fitted (e.g., for historically unprecedented values of independent variables in a regression model).

Confidence intervals: A point forecast should always be accompanied by a confidence interval to indicate the accuracy that is claimed for it, but what does "confidence" mean? It's sort of like "probability," but not exactly. Rather,

- An x% confidence interval is an interval calculated by a *rule* which has the property that the interval will cover the true value x% of the time under *simulated* conditions, *assuming the model is correct*.
- *Loosely speaking*, there is an x% probability that *your* future data will fall in *your* x% confidence interval for the forecast—but only if your model and its underlying assumptions are correct and the sample size is reasonably large. This is why we test model assumptions and why we should be cautious in drawing inferences from small samples.[6]
- If the true distribution of the noise is a *normal* distribution, then a *confidence interval for the forecast* is equal to the point forecast plus-or-minus some number of forecast standard errors, that number being the so-called "critical *t*-value":

[5] A "linear" forecasting model is one in which the forecast is a linear function of other variables whose values are known. The mean model is the simplest case (a trivial example of a linear function), and linear regression models and ARIMA models are more general cases. In the case of the mean model, the parameter risk is a constant, the same for all forecasts. In more general models, the parameter risk associated with a particular forecast depends on the values of independent variables that are multiplied by the parameters. The parameter risk is larger for values of the independent variables that are extreme relative to the values in the sample of data to which the model was fitted.

[6] If the sample size is small, then information that you possess "prior" to the data analysis—even if it is very subjective—becomes relatively much more important and should be taken into account when making inferences or predictions. So-called "Bayesian" methods of inference and prediction provide a systematic way for doing this and are increasingly being used throughout the field of statistics. The statistics department at Duke University is one of the world's leading centers of research in this area.

Confidence interval = forecast ± (critical *t*-value) × (standard error of forecast)

- More precisely, for a confidence interval with confidence level *p*, the appropriate number of standard errors is the "critical value of the *t* distribution with a tail area probability of 1–*p* and *d* degrees of freedom for error", where the number of degrees of freedom ("d.f.") is the sample size (*n*) minus the number of parameters which have been estimated from it (which is 1 in the case of the mean model).

- In Excel, the critical *t*-value for a 2-sided confidence interval with confidence level *p* when there are *d* degrees of freedom is given by the formula T.INV.2T(1–*p*, *d*), or just TINV(1–*p*, *d*) in older versions of Excel.

- When the 1-parameter mean model is fitted to our 20-observation sample of data, the number of degrees of freedom is 20 – 1 = 19, so for a 95% 2-sided confidence interval, the critical *t*-value is **T.INV.2T(5%, 19)** (or just TINV(5%,19) in older versions of Excel), which comes out to be **2.093**.

Here is a so-called "*t*-table" showing the critical values of the *t* distribution for some representative values of the confidence level and the number of degrees of freedom:

Confidence level (for 2-sided confidence interval):

# d.f.:	50%	68%	80%	90%	**95%**	99%	99.7%
Infinity	**0.67**	**1**	**1.28**	**1.64**	**1.96**	**2.58**	**2.97**
200	0.68	1	1.29	1.65	**1.97**	2.60	3
100	0.68	1	1.29	1.66	**1.98**	2.63	3.04
50	0.68	1	1.30	1.68	**2.01**	2.68	3.12
20	0.69	1.02	1.33	1.73	**2.09**	2.85	3.4
10	0.70	1.05	1.38	1.81	**2.23**	3.17	4.02
Rule-of-thumb value:	**2/3**	**1**	**4/3**	**5/3**	**2**	**8/3**	**3**

As the number of degrees of freedom goes to infinity, the *t*-distribution approaches a standard normal distribution, whose critical values are shown in the first row of the table. As you can see, the critical *t*-value is not very sensitive to the number of degrees of freedom except for very low numbers of degrees of freedom in conjunction with high levels of confidence (the shaded cells in the lower right). In most cases the critical values of the *t* distribution are not much different from those of the standard normal distribution. The row below the table shows the "rule of thumb" values that closely approximate the actual critical *t*-values in most situations. In particular, the rule-of-thumb value for a 95% confidence interval is 2, so….

A 95% confidence interval is (roughly) the forecast "plus-or-minus two standard errors."

More rules of thumb for confidence intervals:

- For $n>>20$, a 68% confidence interval is roughly plus-or-minus *one* standard error, a 95% confidence interval is plus-or-minus *two* standard errors, and a 99.7% confidence interval is plus-or-minus *three* standard errors.
- A 50% confidence interval is roughly plus or minus *two-thirds* of a standard error, which is one-half the width of an 80% confidence interval, one-third the width of a 95% confidence interval, and one-quarter the width of a 99% confidence interval.
- A confidence interval that covers 95% of the data is often *too wide* to be very informative. A 1-out-of-20 chance of falling outside some interval is a bit hard to visualize. 50% (a "coin flip") or 90% (1-out-of-10) might be easier for a non-specialist to understand. Most statistical software, including RegressIt, allows you the option to choose the level of confidence for which to calculate confidence intervals.
- Because the distribution of errors is generally bell-shaped with a central peak and "thin tails" (you hope!), the 95% limits are pretty far out compared to where most of the data is really expected to fall, i.e., they may make the forecast look less accurate than it really is.
- Another thing to consider in deciding which level of confidence to use for constructing a confidence interval is *whether or not you are concerned about extreme events*. If it is a high-stakes decision, you may be very interested in the low-probability events in the tails of the distribution, in which case you might really want to focus attention on the location of the edge of the 95% or even the 99% confidence interval for your prediction. (In this case you will also be very interested in whether the distribution of errors is really a normal distribution! If it isn't, these calculations will not be realistic.) But if it is a routine or low-stakes decision, then maybe you are more interested in describing the middle range of the distribution, e.g., the range in which you expect 50% or 80% or 90% of the values to fall.
- My own preference: *just report the forecast and its standard error and leave it to others to apply the rules above to obtain whatever confidence intervals they want.*

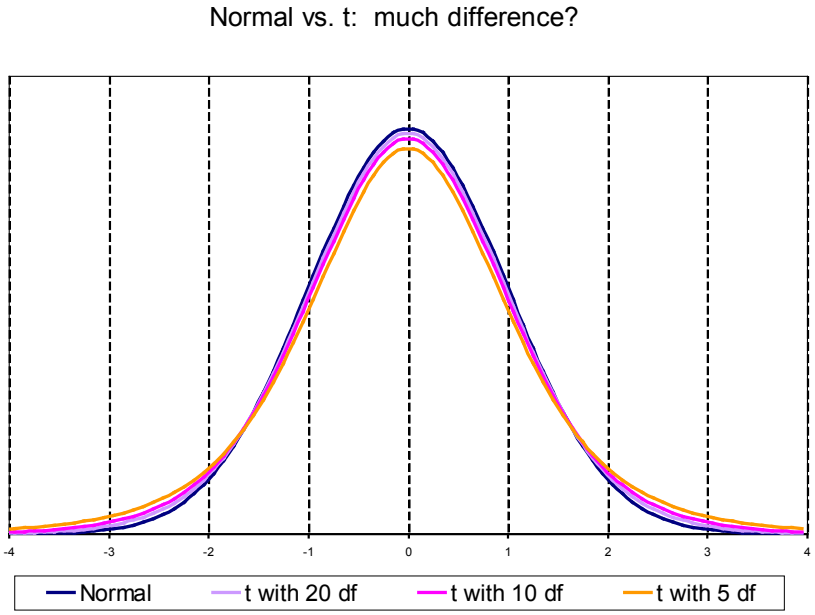
More about *t*:

The *t* distribution is called "Student's *t* distribution" because it was discovered by W.S. Gossett of the Guinness Brewery, who was a pioneer of statistical quality control in the brewing industry and who also published scientific articles anonymously under the pen name "Student". Mathematically, the *t* distribution is the distribution of the quantity

$$(\bar{X} - \mu) \Big/ SE_{mean}$$

which is the *number of standard errors by which the sample mean deviates from the true mean* when the standard deviation of the population is unknown (i.e., when SE_{mean} is calculated from s rather than σ). The *t* distribution resembles a standard normal (z) distribution but with slightly

"fatter tails" when the number of degrees of freedom is small. As pointed out above, the *t*-distribution approaches a normal distribution as the number of degrees of freedom goes to infinity. The following chart shows a comparison of the normal distribution and *t* distribution for 5, 10, and 20 degrees of freedom. As you can see, they are quite close:



In comparing a *t* distribution to a standard normal distribution, what matters is the "tail area probability" that falls outside some given number of standard errors: a 95% confidence interval is the number of standard errors plus-or-minus outside of which there is a tail area probability of 5%. For a lower number of degrees of freedom the "tails" are slightly "fatter", so a greater number of standard errors is needed for a given level of confidence that your estimate won't deviate from the true value by more than that amount. But in most cases, the empirical rules of the thumb given below the *t*-table on page 16 are a very good approximation.

Our example continued:

Time series X (n=20):

114, 126, 123, 112, 68, 116, 50, 108, 163, 79, 67, 98, 131, 83, 56, 109, 81, 61, 90, 92

The *true* mean and standard deviation of the population from which this time series was randomly sampled are $\mu = 100$, $\sigma = 30$, unbeknownst to us in real life. The *sample* mean and standard deviation, which are all we can observe, are:

$$\bar{X} = 96.35 \approx 96, \qquad s = 28.96 \approx 29$$

…and the standard errors of the mean and forecast are:

$$SE_{mean} = 28.96/\sqrt{20} = 6.48 \approx 6.5$$

$$SE_{fcst} = \sqrt{28.96^2 + 6.48^2} = 29.68 \approx 30$$

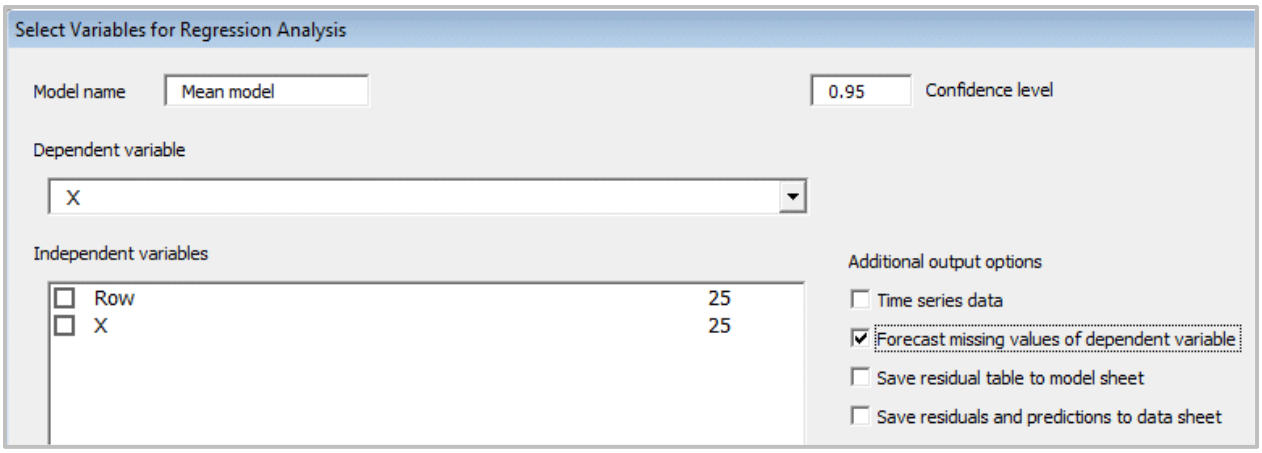
We can use the standard error of the forecast to calculate confidence intervals for the forecasts:

- 95% confidence interval = $96.35 \pm 2.093 \times 29.68 \approx$ **[34, 158]**
- 50% confidence interval = $96.35 \pm 0.688 \times 29.68 \approx$ **[76, 117]**
- These are based on the critical *t*-values T.INV(5%, 19) = 2.093 and T.INV(50%, 19) = 0.688.

Let's suppose that X is stored in a column on a worksheet next to a column of row index numbers (the time scale), and suppose that the time scale extends to row 25, representing future time periods for which forecasts are desired, and that Row and X have been assigned as range names for the two columns:

	A	B
1	Row	X
2	1	114
3	2	126
4	3	123
5	4	112
6	5	68
7	6	116
8	7	50
9	8	108
10	9	163
11	10	79
12	11	67
13	12	98
14	13	131
15	14	83
16	15	56
17	16	109
18	17	81
19	18	61
20	19	90
21	20	92
22	21	
23	22	
24	23	
25	24	
26	25	

If you use RegressIt to fit this model by choosing X as the dependent variable and not specifying any independent variable(s), and if you use the default 95% level for confidence intervals, and check the box for forecasting missing values of the dependent variable, and give it the name "Mean model", like this:



…then the parameter estimates and forecasts come out looking like this:

Model: Mean model
Dependent Variable: X
Equation:
Predicted X = 96.350

Regression Statistics: Mean model for X (0 variables, n=20)

R-Squared	Adj.RSqr	Std.Err.Reg.	# Cases	# Missing	t(2.50%,19)	Conf. level
0.000	0.000	28.963	20	0	2.093	95.0%

Summary Table: Mean model for X (0 variables, n=20)

Variable	Coefficient	Std.Err.	t-Stat.	P-value	Lower95%	Upper95%
Intercept	96.350	6.476	14.877	0.000	82.795	109.905

Forecasts: Mean model for X (0 variables, n=20)

Obs#	Forecast	StErrFcst	Lower95%F	Upper95%F	StErrMean	Lower95%M	Upper95%M
21	96.350	29.679	34.232	158.468	6.476	82.795	109.905
22	96.350	29.679	34.232	158.468	6.476	82.795	109.905
23	96.350	29.679	34.232	158.468	6.476	82.795	109.905
24	96.350	29.679	34.232	158.468	6.476	82.795	109.905
25	96.350	29.679	34.232	158.468	6.476	82.795	109.905

Don't worry that this model has an R-squared of zero! This is always the case in a mean model (i.e., an intercept-only model). *The important number is the standard error of the model (labeled here as "standard error of the regression"), whose value is 28.96, which measures the standard deviation of the variations in the past data that have not been explained by the model.*

Note that each separate table (and graph) in RegressIt output is tagged with the model name and the sample size. This may look redundant, but it can be very important in practice. Often a lot of models are fitted to the same variables during one or more analysis sessions, and pieces of output

from the best model(s) are later copied and pasted into reports. Unless more detailed titles are added by hand at that stage, it is not always clear which model produced the displayed results.

Presentation of results:

In presenting your results to others, *know what to round off and when* in the number of decimal places you choose to display.[7] Don't overwhelm your audience with more digits than are meaningful for understanding the results. (Alas, many widely used statistical software packages routinely report 10 or more significant digits of precision in their output for any procedure, and the numbers are often poorly formatted with misaligned decimal points.) The calculations shown above were rounded off to integer values, and it is uninformative to report any more decimal places than that. It would be silly to report the 95% interval as, say, [34.232, 158.468], or even as [34.2, 158.4].[8] The forecast standard error is 30, so rounding off to integer values is precise to within 1/60 of a standard error. Just because your software can give you up to 15 digits of precision in any calculation doesn't mean that you need to display all of them! It would be simplest in this case to just say that **the forecast is 96 with a standard error of 30.**

In general it is good to choose units for your variables that are mutually consistent (e.g., don't measure some variables in thousands of dollars and others in the same model in billions of dollars) and which yield coefficient estimates that don't differ wildly in orders of magnitude and don't have too many digits before the decimal place or too many zeros after it.

Our example continued:

Under the assumptions of the mean model, the forecast and confidence interval for the next period also apply to all future periods, as shown in the forecast table above. The model assumes that all future observations will be drawn from the same distribution, from here to eternity. So, we can extrapolate this one forecast arbitrarily far into the future, and a plot of the forecasts and confidence intervals stretching off into the future will look like a set of *parallel horizontal lines*.

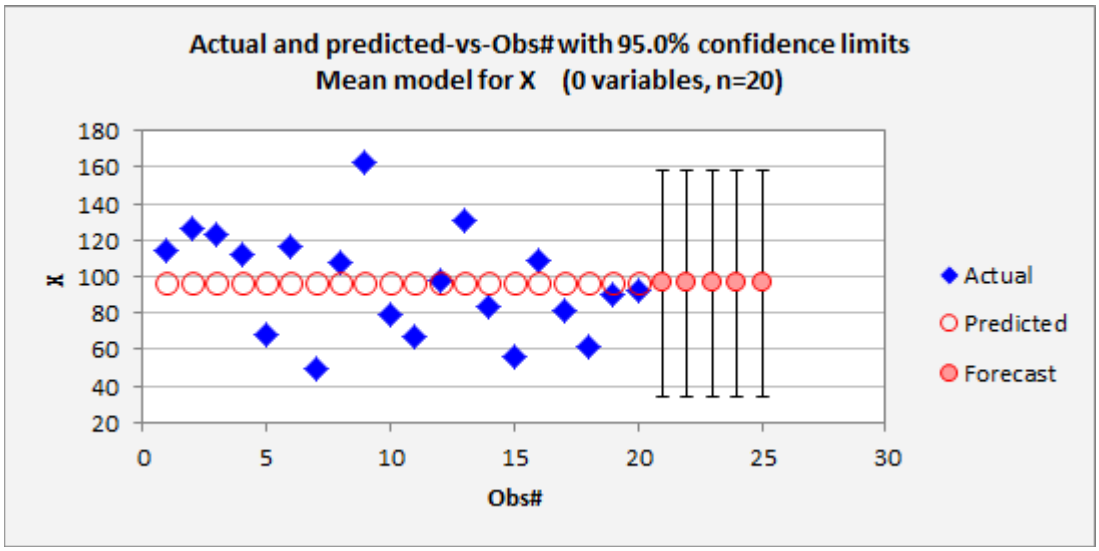
If X is a time series, such as a series of weekly or monthly observations of an economic variable or a weather variable, the long-horizon forecasts produced by the mean model do not look very realistic. The distant future is usually more uncertain than the immediate future, and most time series forecasting models will reflect this by yielding confidence intervals that gradually get wider at longer horizons. The longer-horizon forecasts for the mean model make more sense if X is not really a time series, but rather a set of independent random samples from a population that is not changing with time, in which case all the observations are statistically the same whether they are performed today or tomorrow, sequentially or in bunches.

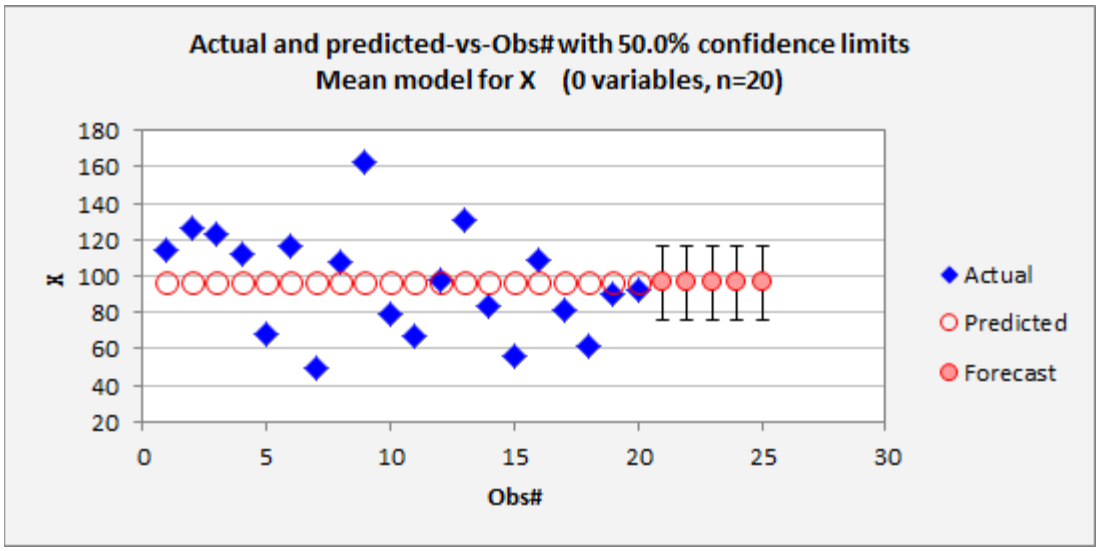
---

[7] You can use the "increase decimal" and "decrease decimal" functions on the Home toolbar in Excel to adjust the number of displayed decimals, and the number displayed in Excel will be the default number displayed in Word or Powerpoint if you copy a table of numbers.

[8] About the only reason for reporting your results with more digits than are really significant is for audit-trail purposes in case someone tries to exactly reproduce your results. But for that purpose you can just keep your spreadsheet files. In fact, you should ALWAYS be sure to keep your spreadsheet files and make sure that the "final" version has a file name that clearly identifies it with your final report.

Here are plots of the forecasts and confidence intervals produced by RegressIt for periods 21 through 25 at the 95% level of confidence and the 50% level of confidence:

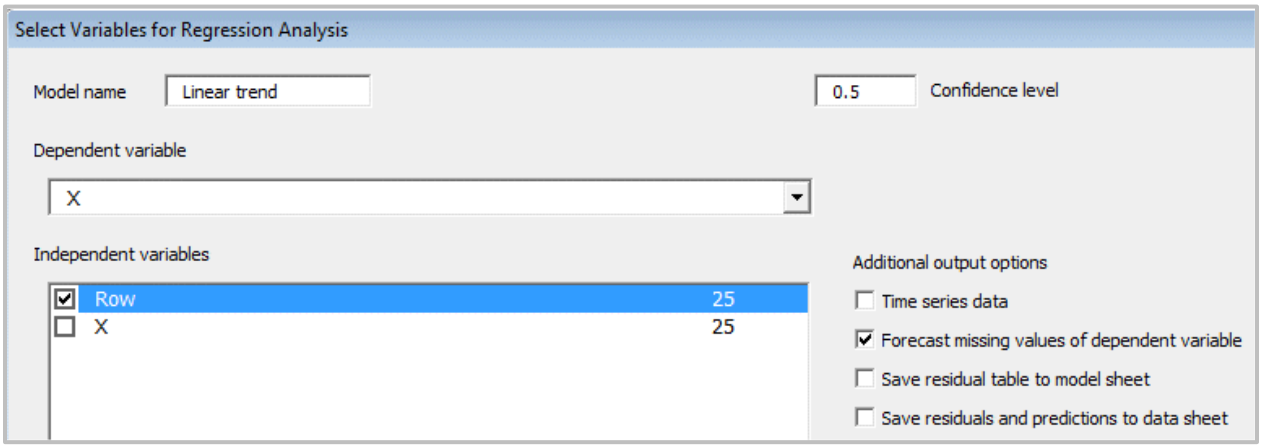




Note that the 50% confidence interval is about exactly 1/3 the width of the 95% confidence interval. And, as pointed out above, the width of the confidence intervals for forecasts from the mean model remains constant as we forecast farther into the future, because all periods in the future are expected to be the statistically identical to those in the past.

The calculation of confidence intervals shown above takes into account the *intrinsic risk* and *parameter risk* in the forecast, assuming the model is correct, i.e., assuming the mean and standard deviation really are constant over time, and assuming that variations from the mean really are statistically independent from one period to another.

But what about model risk? For example, suppose it is possible that there is some kind of trend? In that case it might be appropriate (as a first-order approximation) to fit a *linear trend model*, which is a simple regression model in which the independent variable is the row index or any ascending sequence of equally spaced numbers. The data set here contains a variable called Row that can be used for this purpose. Here is how the linear trend model would be specified in RegressIt, with the initial confidence level[9] set to 50%:



…and here is the summary output:

Model: Linear trend model
Dependent Variable: X
Equation:
Predicted X = 114.611 - 1.739*Row

Regression Statistics: Linear trend model for X (1 variable, n=20)

R-Squared	Adj.RSqr	Std.Err.Reg.	# Cases	# Missing	t(25.00%,18)	Conf. level
0.126	0.078	27.816	20	0	0.688	50.0%

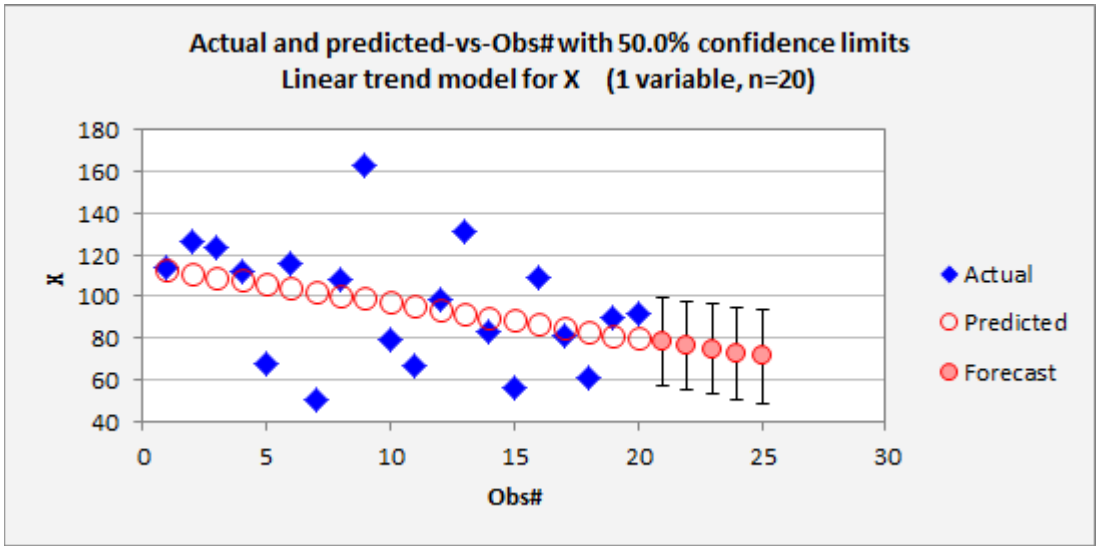
Summary Table: Linear trend model for X (1 variable, n=20)

Variable	Coefficient	Std.Err.	t-Stat.	P-value	Lower50%	Upper50%
Intercept	114.611	12.921	8.870	0.000	105.716	123.505
Row	-1.739	1.079	-1.612	0.124	-2.482	-0.997

This is a different modeling assumption, and it leads to different forecasts and confidence intervals for the next 5 periods, as shown in the table and chart below. *The forecasts now trend downward by 1.739 per period, which is the estimated coefficient of the row-number variable.* The 50% confidence intervals are only slightly wider than those of the mean model—confidence intervals for a linear trend model get wider at longer forecast horizons, but not by much. Notice that the upper 50% confidence limits for this model's forecasts for periods 23-24-25 are all below the point forecast of the first model, which was 96.35.

[9] The confidence level that is specified at the time that the regression is run can be changed on the model output spreadsheet after it has been generated, and the confidence interval calculations and plots will be dynamically updated . This is all that was needed to generate the two different charts shown above.

Forecasts: Linear trend model for X (1 variable, n=20)								
Obs#	Forecast	StErrFcst	Lower50%F	Upper50%F	StErrMean	Lower50%M	Upper50%M	Row
21	78.089	30.671	56.977	99.202	12.921	69.195	86.984	21
22	76.350	31.085	54.952	97.748	13.877	66.798	85.903	22
23	74.611	31.531	52.906	96.316	14.849	64.390	84.833	23
24	72.872	32.007	50.839	94.905	15.835	61.972	83.772	24
25	71.133	32.512	48.753	93.513	16.832	59.547	82.720	25



The Excel file with both of these models is available here:
people.duke.edu/~rnau/Examples_of_mean_and_linear_trend_models.xlsx
Some of the additional rows of output at the top (residual distribution statistics and ANOVA table) have been hidden for simplicity.

Is the linear trend model a plausible alternative model, based on what the data is telling us? Trend line models will be discussed in more detail later, but for a preview of that discussion, one indicator we can look at is the *statistical significance of the estimated trend coefficient* in the regression model, as measured by its *t*-statistic. The *t*-statistic of a coefficient is its estimated value divided by its own standard error, which is its "number of standard errors from zero." The rule-of-thumb standards are that a coefficient estimate is significant at the 0.05 level if its *t*-stat is greater than 2 in magnitude and it is significant at the 0.10 level if it is greater than 1.67 in magnitude. The *t*-stat of the trend coefficient is 1.612 in magnitude in this case, as seen in the regression summary table above, which is not significant even at the 0.10 level. (Its exact level of significance, which is its P-value, is 0.124.) This is not so insignificant as to be completely meaningless but probably not big enough to be useful for decision making unless it is accompanied by other evidence in favor of a negative trend. Another thing that most naïve individuals immediately ask about a regression model is: "what's the R-squared?" The (adjusted) R-squared is the percentage by which the estimated error variance of the regression model (the square of the standard error of the regression) is less than that of the mean model, i.e., it is the percent of variance that the model thinks it has "explained." In this case it is 7.8%, which is not very impressive.

A better thing to look at is the *standard error of the regression* (another name for the standard error of the model), which is a lower bound on its forecast standard error. The standard error of the regression is 27.82 in this case, compared to a standard error of 28.96 for the mean model, which is not a big improvement—in fact, it is only a 4% reduction!—so it's hard to choose on that basis.

> A rule of thumb: when adjusted R-squared is fairly small (say, less than 20%), the percentage by which the standard error of the regression model is less than the standard error of the mean model is roughly one-half of adjusted R-squared.

More importantly, *these two numbers cannot both be correct as far as errors in predicting the future are concerned, because each one is based on the assumption that its own model is the correct one, and the models are very different*. If the mean model is correct in its assumptions, then the standard error of the linear trend model is a very unrealistic estimate of the accuracy of its forecasts for the future, and vice versa. The model that *thinks* it is best is not always right, which is why *you* need to exercise your own judgment.

So, the data does not strongly indicate that there is a trend. However, this is a small sample, so we shouldn't rely only on the data for such an important inference. We should take into account everything we know about the situation that might indicate whether a linear trend is to be expected, and if so, whether the estimated trend is at least in the right direction. But if we *do* choose the linear trend model, it would be dangerous to extrapolate the trend very far into the future based on such a small sample, because *the longer-horizon forecasts and confidence intervals depend very sensitively on the assumption that the trend is linear and constant.*

A linear trend model is not a very "robust" model for time-series forecasting. If you have no a priori knowledge of whether the series has a positive trend, negative trend, or zero trend, then it is more conservative to assume zero trend (i.e., to stick with the mean model or perhaps a *moving average* model that puts more weight on the most recent values) than to use a linear trend model with a not-very-significant trend estimate. However, if you have good reasons for believing there *is* a trend in a particular direction, but you just don't know if it is steep enough to stand out against the background noise in a small sample, then the linear trend model might be logically preferred even if its estimated trend coefficient is only marginally significant. Still, its confidence intervals for *long-horizon* forecasts probably shouldn't be taken very seriously.

Another problem with the linear trend model is that in fitting the trend line it gives equal weight to all the data, old and new. It tries just as hard to fit the very first data points as the very last ones. In real applications, you are generally more interested in how the model has been doing lately, i.e., what is happening at the "business end" of the time series. Alternative models for trend extrapolation, which place relatively more weight on the most recent data, will be discussed in later sections of these notes.

There are also other ways in which the mean model might turn out to be the wrong model for a given data set. For example:

- Consecutive deviations from the mean might be correlated with each other
- The standard deviation of variations around the mean might not be constant over time
- The mean may have undergone a "step change" at some point in the middle of the series

Each of these alternative hypotheses could be tested by fitting a more complicated model and evaluating the significance of its additional coefficients, but once again, you should not rely only on significance tests to judge the correctness of the model—you should draw on any other knowledge you may have, particularly given the small sample size. You shouldn't just blindly test a lot of other models without good motivation.

At the end of the day you may have to explain your choice of model to other decision makers, perhaps betting their money and your reputation on it, and you ought to have a better argument than just "this one has good t-stats" or "it has an R-squared of 85%" or (worst of all) "my automatic forecasting software picked it for me."

The mean model is very simple, but it is the foundation for more sophisticated models we will encounter later: regression, random walk, and ARIMA. It has the same generic features as any statistical forecasting model:

→ One or more *parameters* to be estimated

→ An estimate of the *intrinsic* risk, which is called the "standard error of the model"

→ An estimate of the *parameter* risk(s), which is called the "standard error of the coefficient(s)"

→ A *forecast standard error* that reflects both intrinsic risk & parameter risk…

…AND MODEL RISK TOO!

The formulas for calculating standard errors and confidence limits when using the mean model have exactly the same form as those that apply to regression models. If you understand the former, you will understand the latter. The only difference is that *in a regression model the estimated mean and its standard error are not constants—they depend on the values of the independent variables*—and so the standard errors of forecasts are not all the same either. The standard errors of forecasts are larger, and hence confidence intervals are wider, for predictions that are made for more extreme conditions or more distant points in the future.

To sum up the more general lessons of this chapter:

There are no **magic formulas** for creating good forecasting models, but there are **best practices** you should try to learn, and that's what this web site is about.

t-stats, P-values, and R-squared, and other test statistics are numbers you should know how to interpret and use, but they are not the most important numbers in your analysis and they are not the bottom line.

Your bottom-line issues are the following:

- First of all, *what new things have you learned* from your data collection and from your analysis that might be useful to yourself or others? Perhaps you learned something new about the amount of uncertainty in Y or about trends in Y or about the relation between Y and X, and perhaps you learned that you need better quality data or more sophisticated models or software in order to be able answer questions like that. Even the latter lesson may be very useful, if you act upon it!
- *What assumptions does your model make*, and do they seem reasonable based on everything you know?
- *Would these assumptions make sense to someone else,* and can you explain them in plain language?
- Would a simpler, easier-to-understand model perform almost as well?
- *How accurate are your model's predictions for the sample data in real terms*, in comparison to those of alternative models?
- How accurate is it likely to be when you use it to *predict the future* (not merely to fit what has been observed in the past)?
- How good (or bad) are the *inferences* and *decisions* you will make when using it?