

### Homework Set #1 (10 points)

You will use the following datasets to investigate the trend in a hypothetical invasive species. Your task is to determine the rate (individuals / year) of its increase in relative abundance.

The study starts at time 0 – the year before the species was detected for the first time – and runs for 50 years. You will use two datasets: a fine-scale study (of surveys collected every year) and a coarse-scale study (of surveys collected every 5 years).

In addition, you will consider three sampling methods, with different built-in errors:

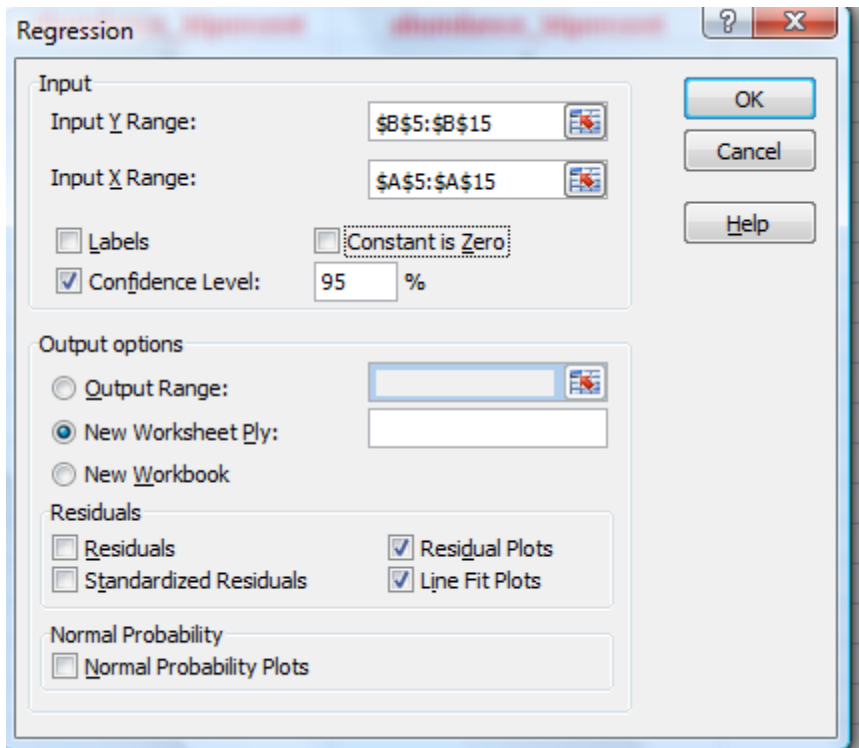
- One method has no error (“abundance\_noerror”)
- One method has a 10% error (“abundance\_10percent”)
- One method has a 50% error (“abundance\_50percent”)

So, you have 6 time series: 3 short and 3 long.

For each one, you will calculate a regression, and will provide the best-fit regression equation

Note: Make sure you select the following input / output parameters:

- 95 % Confidence Intervals
- Residual Plots, Line fit Plots



1) Report the following statistics for each analysis, indicating the time series length / error rate:  
(+1 point for each regression)

- **For instance: coarse time series, no error**

### SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.999969
R Square	0.999937
Adjusted R Square	0.999931
Standard Error	0.275241
Observations	11

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	10900.23	10900.23	143883	3.13E-20
Residual	9	0.681818	0.075758		
Total	10	10900.91			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.318182	0.155257	2.04939	0.070683	-0.03303	0.669397
X Variable 1	1.990909	0.005249	379.3191	3.13E-20	1.979036	2.002782

Copy and paste the plots of the best-fit trend and the residuals, for each regression you perform.

Answer these questions:

- Were the trends (number per year) that you documented for the six examples significant? (refer to the p values and to the slope coefficient +/- 95 confidence intervals) (+2 points)
- How well do these models describe the population trend? (what is the proportion of the variance explained by the linear trend model?) (+2 points)
- Describe what analyses gave you the highest degree of certainty about the observed trend? (explain why, using the answers above) (+2 points)

**1a) Fine-scale, No error:** This model has an almost perfect fit (explains 99.99% of the variance), and detected a significant ( $p = 0.049$ ) slope, whose 95% confidence intervals (CI) (1.995 – 2.000) overlap the real slope ( $b = 2$ ). The plot of the line and the residuals, underscore the great fit of this model without any noise (no error). The CV of the estimated slope (the *XVariable1* coefficient) is fairly small (0.357).

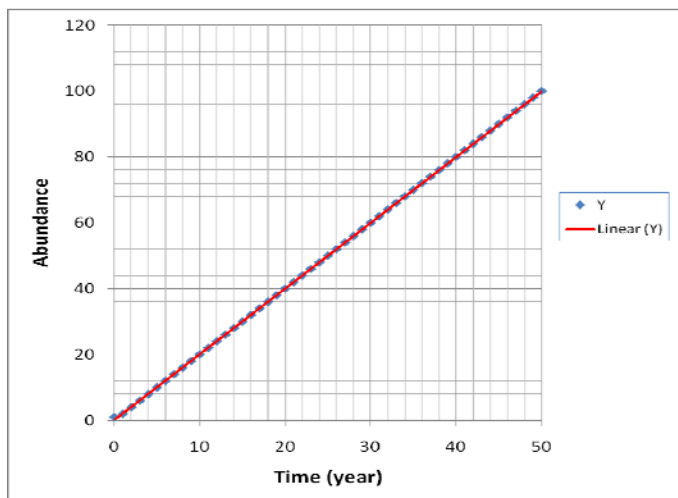
## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999
R Square	0.999
Adj R Square	0.999
Standard Error	0.137
Observations	51

## ANOVA

	df	SS	MS	F	Significance F
Regression	1	44100.060	44100.060	2339067	2.6E-116
Residual	49	0.923	0.018		
Total	50	44100.980			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.076	0.037	2.009	<b>0.049</b>	1.52E-05	0.152
X Variable 1	1.997	0.001	1529.401	2.6E-116	1.995	2.000



**1b) Fine-scale, 10% error:** This model has an almost perfect fit (explains 99.99% of the variance), and detected a significant ( $p < 0.05$ ) slope, whose 95% confidence intervals (CI) (1.987 – 2.011) overlap the real slope ( $b = 2$ ). The plot of the line and the residuals, underscore the great fit of this model with some noise (10% error). The CV of the estimated slope (the *XVariable1* coefficient) is larger (1.786) than in the model with no error.

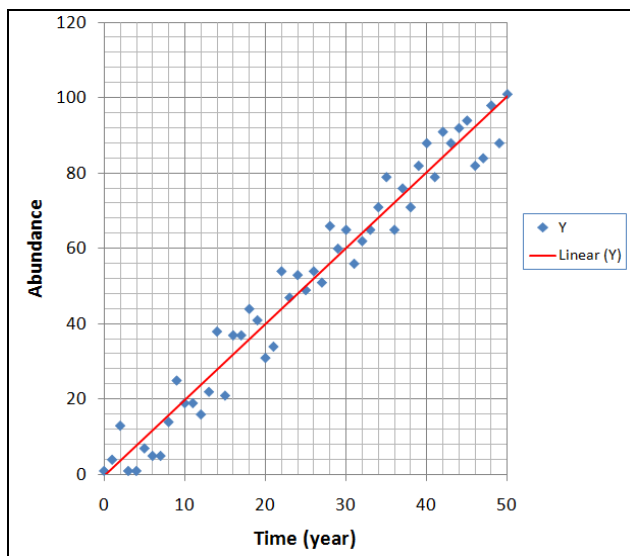
## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999
R Square	0.999
Adj. R Square	0.999
Standard Error	0.627
Observations	51

## ANOVA

	df	SS	MS	F	Significance	
					F	
Regression	1	44182	44182	112152	5.49E-84	
Residual	49	19.303	0.393944			
Total	50	44201.31				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.019	0.173	0.115	0.908	-0.328	0.368
<i>X Variable 1</i>	1.999	0.005	334.892	<b>5.49E-84</b>	1.987	2.011



**Ic) Fine-scale, 50% error:** This model has an almost perfect fit (explains 80.6% of the variance), and detected a significant ( $p < 0.05$ ) slope, whose 95% confidence intervals (CI) (1.626 – 2.152) overlap the real slope ( $b = 2$ ). The plot of the line and the residuals, showcase the poor fit of this model with a great deal of noise (50% error). The CV of the estimated slope (the XVariable1 coefficient) is larger (49.146) than in the model with no error.

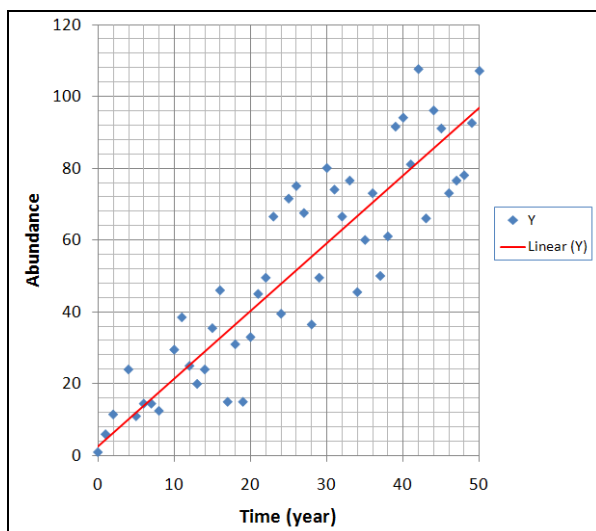
## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.899
R Square	0.809
Adj. R Square	0.806
Standard Error	13.745
Observations	51

## ANOVA

	df	SS	MS	F	Significance F
Regression	1	39458.480	39458.480	208.851	2.69E-19
Residual	49	9257.612	188.9309		
Total	50	48716.090			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.463	3.793	0.649	0.519	-5.159	10.087
X Variable 1	1.889	0.130	14.451	<b>2.69E-19</b>	1.626	2.152



**Id) Coarse-scale, No error:** Almost perfect fit (model explains 99.99% of the variance) and significant slope ( $p < 0.05$ ). The point estimate has very low variability, as evidenced by the small CV (0.833).

The plot of the residuals and small standard error of the regression results underscore the great fit of this model, which did not include any noise (no error). Thus, we detected the trend (slope = 2) very well in this dataset.

#### SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999
R Square	0.999
Adj. R Square	0.999
Standard Error	0.275
Observations	11

#### ANOVA

	df	SS	MS	F	Significance F
Regression	1	10900.230	10900.230	143883	3.13E-20
Residual	9	0.681	0.075		
Total	10	10900.910			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.318	0.155	2.049	0.070	-0.033	0.669
X Variable 1	1.990	0.005	379.319	<b>3.13E-20</b>	1.979	2.002

**1e) Coarse-scale, 10% error:** Very good fit (explains 99.99% of the variance) and significant slope ( $p < 0.05$ ). The 95% CI of the best-fit slope (1.979 – 2.002) includes the real value ( $m = 2$ ), and the point estimate (1.986) is fairly precise but more variable ( $CV = 2.004$ ). Yet, the plot of the residuals and the larger standard error of the regression results underscores that the added noise (10% error) is blurring our ability to detect the trend (slope = 2).

## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999
R Square	0.999
Adj R Square	0.999
Standard Error	0.663
Observations	11

## ANOVA

	df	SS	MS	F	Significance F
Regression	1	10848.530	10848.530	24642.160	8.78E-17
Residual	9	3.962	0.440		
Total	10	10852.490			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.290	0.374	0.777	0.456	-0.555	1.137
X Variable 1	1.986	0.012	156.978	<b>8.78E-17</b>	1.957	2.014

*If) Coarse-scale 50% error: Lower fit (explains 94.50% of the variance) and significant slope ( $p < 0.05$ ). While the 95% CI of the best-fit slope (1.363 – 2.327) overlaps the real value ( $b = 2$ ), it is much more variable than the previous two model results ( $CV = 38.109$ ). The point estimate (1.845) is considerably less precise than the results of the first two models.*

## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.945
R Square	0.893
Adj R Square	0.881
Standard Error	11.164
Observations	11

## ANOVA

	df	SS	MS	F	Significance F
Regression	1	9365.682	9365.682	75.144	1.16E-05
Residual	9	1121.727	124.636		
Total	10	10487.410			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.954	6.297	0.627	0.545	-10.291	18.200
X Variable 1	1.845	0.212	8.668	<b>1.16E-05</b>	1.363	2.327

**TAKE HOME LESSON:**

*The 'true trend' of these data was a linear increase in abundance ('true' slope = 2.00), despite an initial oscillation when the abundance increased from 1 individual in year 0 to 2 individuals in year 1. Thereafter, the populations increased by the same amount each year (2 individuals).*

*Overall, the analysis with the fine scale data without measurement error described the actual trend most accurately (small confidence interval, high adjusted  $R^2$ , very accurate slope coefficient, low SE). However, small amounts of error (10%) still provided significant results with a slope coefficient estimate that overlapped the true slope. High degrees of noise (error = 50%) obscured the results, whether the fine-scale or coarse-scale data were used in the analysis.*

- 2) Next, you will explore how the record length (the number of samples) influences your ability to detect a significant increase.

To do this, you will use all six time series and – starting with the first three samples (times 0, 1, 2 for the fine-scale data OR times 0, 5, 10 for the coarse-scale data), you will perform regressions increasing your sample size by 1 sample at a time (Hint: you will repeat the same regression with 3, 4, 5, 6, ... samples). STOP the first time you get a significant increase for each time series.

For each step, report the p value and the slope (coefficient +/- 95 % confidence intervals). (+1 point for each dataset: fine\_scale and coarse\_scale)

Answer these questions:

- **What was the shortest duration you had to sample to determine a significant increase? What time series were you analyzing? (+1 point)**

*The fine-scale time series with no error yielded a significant result on the first year of the analysis (year 3) and subsequently. Yet, the slope estimate improved (smaller CV and 95% CI), as we added more years of data to the time series. The time series with 10% error required four data points to detect a significant trend (detected at year 4). The time series with 50% error required 13 years of sampling before a significant increase was detected. Furthermore, this highly-variable dataset highlights the problem of detecting the wrong trend in small time series Note that the coefficient (slope) resulting from this analysis varied widely, and was first significant after only 3 years of data; yet, the estimate did not overlap the real slope ( $b = 2$ ) (See entry in blue font on Table 3).*

- **What was the longest duration you had to sample to determine a significant increase? What time series were you analyzing? (+ 1 point)**

*I analyzed the fine-scale time series with no error, 10% error and 50% error.*

**Table 1:** Results of the incremental analysis for the fine-scale time series without error (year 3 to 13), showing the *p* values (red font indicates significant results at the  $\alpha = 0.05$  level), the best-fit coefficient (slope) and the SE and 95% CI of that point estimate.

<b>treatment</b>	<b>n</b>	<b>p</b>	<b>coeff</b>	<b>SE</b>	<b>lower</b>	<b>upper</b>
noerror	3	0.017498	1.5	0.144338	0.416017	3.083983
noerror	4	0.010222	1.7	0.173205	0.954759	2.445241
noerror	5	0.000574	1.8	0.11547	1.432523	2.167477
noerror	6	2.3E-05	1.857143	0.082479	1.628146	2.08614
noerror	7	6.99E-07	1.892857	0.061859	1.733844	2.051871
noerror	8	1.67E-08	1.916667	0.048113	1.79894	2.034394
noerror	9	3.25E-10	1.933333	0.03849	1.842319	2.024348
noerror	10	5.24E-12	1.945455	0.031492	1.872834	2.018075
noerror	11	7.17E-14	1.954545	0.026243	1.895179	2.013912
noerror	12	8.46E-16	1.961538	0.022206	1.912061	2.011016
noerror	13	8.7E-18	1.967033	0.019034	1.92514	2.008925

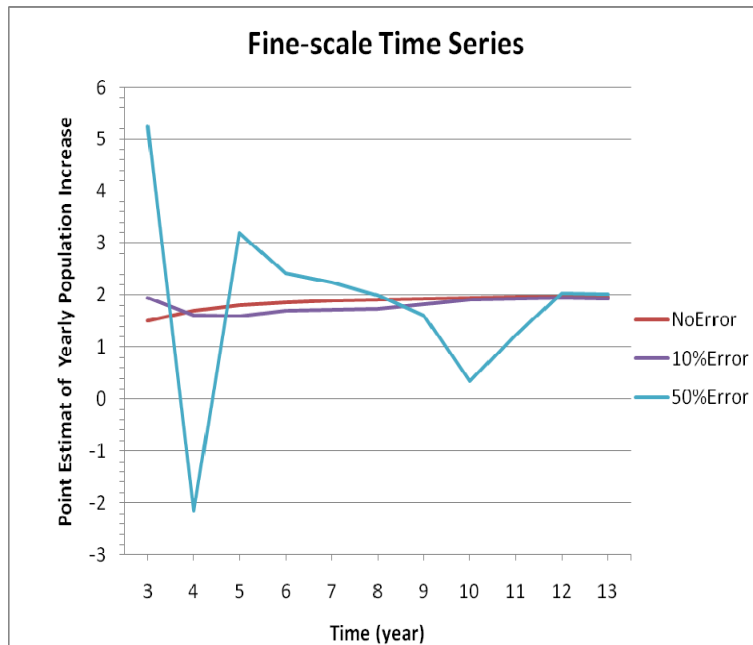
**Table 2:** Results of the incremental analysis for the fine-scale time series with 10% error (year 3 to 13), showing the *p* values (red font indicates significant results at the  $\alpha = 0.05$  level), the best-fit coefficient (slope) and the SE and 95% CI of that point estimate.

<b>treatment</b>	<b>n</b>	<b>p</b>	<b>coeff</b>	<b>SE</b>	<b>lower</b>	<b>upper</b>
10%error	3	0.139109	1.95	0.433013	-3.55195	7.451948
10%error	4	0.026988	1.62	0.271662	0.451135	2.788865
10%error	5	0.002082	1.59	0.157797	1.087818	2.092182
10%error	6	0.000147	1.697143	0.120407	1.362838	2.031448
10%error	7	5.91E-06	1.725	0.086647	1.502268	1.947732
10%error	8	1.84E-07	1.742857	0.0654	1.582828	1.902886
10%error	9	2.06E-08	1.815	0.065583	1.65992	1.97008
10%error	10	6.83E-09	1.910909	0.076179	1.735239	2.086579
10%error	11	2.13E-10	1.930909	0.063263	1.787798	2.07402
10%error	12	5.35E-12	1.937413	0.052772	1.81983	2.054995
10%error	13	1.38E-13	1.926923	0.04501	1.827858	2.025988

**Table 3:** Results of the incremental analysis for the fine-scale time series with 50% error (year 3 to 13), showing the  $p$  values (red font indicates significant results at the  $\alpha = 0.05$  level), the best-fit coefficient (slope) and the SE and 95% CI of that point estimate. The blue font highlights an erroneous significant trend (slope = 5.25 +/- 0.144 SE) obtained after only three years of data; which does not overlap the real trend ( $b = 2$ ).

treatment	n	p	coeff	SE	lower	upper
50%error	3	<b>0.017498</b>	5.25	0.144338	3.416017	7.083983
50%error	4	0.664788	-2.15	4.27288	-20.5347	16.23472
50%error	5	0.477475	3.2	3.953058	-9.38039	15.78039
50%error	6	0.410141	2.414286	2.627342	-4.88038	9.708956
50%error	7	0.280524	2.25	1.860231	-2.53188	7.031876
50%error	8	0.202645	1.994048	1.394387	-1.41789	5.405989
50%error	9	0.18809	1.608333	1.102806	-0.99939	4.216055
50%error	10	0.77227	0.342424	1.143726	-2.29501	2.97986
50%error	11	0.278627	1.227273	1.064452	-1.18068	3.63523
50%error	12	0.069357	2.036713	1.00141	-0.19457	4.267995
50%error	13	<b>0.036563</b>	2.013736	0.84645	0.150712	3.876761

**Figure 4.** Results of the incremental analysis of the time series with different levels of noise (error), showing how the point estimate of the slope approaches the real value ( $b = 2$ ) in all three scenarios, but requires different sample sizes to do so. In particular, the highly-variable estimates from the noisy data (50% error) vary widely and do not converge to the correct answer until year 12.



**TAKE HOME LESSON:** *The ability to detect a trend depends proportionally on the magnitude of the signal and inversely on the degree of noise (variability) inherent in the measurements. For set levels of signal to noise ratios, longer time series will have a higher probability of detecting a trend. Moreover, the longer the time series, the more precise the parameter estimates will be (smaller CV and 95% CIs). While all models may converge to the same parameter estimate eventually, time is precious in conservation. Thus, detecting a trend or a pattern early is critical to implement actions in a timely fashion.*

*Infrequent (coarse-scale) time series will add additional delays in the detection of a trend, by spreading out the sampling over a longer time frame. For instance, given a level of measurement error and a real increasing trend, the 5-year sampling regime will take considerably longer to capture the pattern. First of all, it will take at least 15 years to collect the 3 data points needed to run the first regression. Furthermore, in cases with a lot of noise, it will take a while to smooth out any large measurement errors in the initial values (which anchor the time series). This will require additional samples, collected at 5-year intervals.*

- 3) Finally, consider if the rate of increase had been larger (4 animals per year) and smaller (0.5 animals per year). How do you think your results in section 2 would have changed –would it have taken more or less samples to find a significant increase? (+1 point)

*The larger the signal (the bigger the slope), the easier it will be to detect a significant increase, given any specific level of sampling error. The time series with a slope of 4 would yield significant results sooner (require smaller sample size), while the time series with a slope of 0.5 would yield significant results later (require larger sample size).*

More specifically, discuss how you would expect the power of the analysis (the ability to detect an increase) to be influenced by the following: error in sampling, the actual rate of population increase, and the number of samples taken. Describe whether each factor would increase or decrease the power (+1 point for each).

*I refer you to the following paper (which is available with this key):*

***Barbara L. Taylor and Tim Gerrodette. 1993. The Uses of Statistical Power in Conservation Biology: The Vaquita and Northern Spotted Owl. Conservation Biology, 7: 489-500.***

*The power (or ability of a test to detect a trend) is an increasing function of the Effect Size (the signal, ES) and a decreasing function of the test statistic (the bigger the alpha level, the bigger the power). The power is inversely related to the Variability of the data (the noise, V).*

$$\text{Power} = f\left(\frac{ES - T_{1-\alpha}}{V}\right)$$