**Homework Set #1 (Points for grabs: 25 – Total HW value: 5% of class grade)**

Email to khyrenba@gmail.com with email subject “MARS6400 Hw#1” By end of Feb 2

You will use six made-up datasets (in to investigate the trend in a hypothetical protected species. Your task is to determine the rate (% population / year) of its decrease in relative abundance.

The study starts at time 0 – the baseline year before the annual surveys started and the relative abundance was 100% – and runs for 50 years. You will consider three sampling methods, with different built-in error rates:

* 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 measured yearly for 50 years: 3 are stable and 3 are declining.

For each one, you will calculate a regression, and will provide the best-fit regression equation using the “Regression” tool in the Data Analysis Excel Add-In.

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

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



1. Characterize each time series using tables / figures (+1 point for each: 6 points total).

For each time series do the following:

* Create a scatterplot, showing the relative abundance at time t (x axis) versus the relative abundance at time t+1 (y axis) (Hint: if you start at the baseline datapoint, there should be 50 points in each scatterplot).
* Create a line graph showing the change in population (dN / dt) for each time step (y axis) versus year (x axis). (Hint: if you start at the baseline datapoint, there should be 50 points in each line graph).
* Report the following: mean abundance, STD of abundance and CV of abundance (Hint: do not include the reference datapoint, so there are 50 data points in each time series)
1. Report the following statistics for each time series, indicating the time series length / error rate (+1 point for each regression)

***1a – 2a) Stable, No error:*** *This model has perfect fit (explains 100 % of the variance), and detected no significant slope, whose 95% confidence intervals (CI) (0.0 – 0.0) overlap the point estimate of the slope for a null hypothesis (b = 0). The plots of the line and the residuals underscore the perfect fit of this model without noise (no error).*

**NOTE:** I worked out the various figures for the first example (stable population with no error).

**Figure 1:** Trend in population size over time for the stable trajectory with no error*.*

**Hint:** You are plotting population size (or % population) as a function of time (year). The population size is the dependent variable (y) and time is the independent variable (x).

**Table 1:** Population size statistics for the stable trajectory with no error.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Sample Size*** | ***Mean*** | ***S.D.*** | ***C.V. (%)*** |
| 50 | 100 | 0 | 0 |



**Figure 2:** Change in population size over time for the stable trajectory with no error*.*

**Hint:** You are plotting the change in numbers (from one time step to the next), and the outcomes can be positive (increase) or negative (decrease). There are 50 steps, starting at time t = 1.



**Figure 3:** Relationship between population numbers for any two successive time steps, for the stable trajectory with no error*.*

**Hint:** The phase diagram relates population numbers at any given time step (t) to population numbers at the following time step (t+1). Thus, 50 points are plotted in this scatterplot diagram.

***NOTE: Because this trend has no error, the linear regression fails to yield a p value.***

***There is just not enough variation. However, we can see that this is the case, because the r-squared term = 1.***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 1 |  |  |  |  |  |  |  |
| R Square | 1 |  |  |  |  |  |  |  |
| Adj R Square | 1 |  |  |  |  |  |  |  |
| Standard Error | 0 |  |  |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Analysis of Variance (ANOVA)  |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 0 | 0 | #NUM! | #NUM! |  |  |  |
| Residual | 49 | 0 | 0 |  |  |  |  |  |
| Total | 50 | 0 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 100 | 0 | 65535 | #NUM! | 100 | 100 |
| year | 0 | 0 | 65535 | #NUM! | 0 | 0 |

***Note:*** Even if the p-value is significant (less than 0.05) – this is an artifact of the lack of variation in the abundance data (y axis). Notice that because the slope = 0, this regression cannot be significant. In summary: there is not a significant result, suggesting there is no trend.

***Hint. Some details to note:***

*Degrees of Freedom: Total DF is n-1. Regression df = 1.*

*Coefficients:*

*Intercept = 100 (point estimate) with standard error = 0 (95% Confidence intervals)*

*Year (this is the slope) = 0 (point estimate) with standard error = 0 (95% Confidence intervals)*

Result: The slope of the line = 0. There is no significant decline in abundance over time

***1b – 2b) Stable, 10% error:*** *This model does not fit the data (r square = 0.002) because the only pattern is random noise (variation). Otherwise, there is no linear trend in abundance over time. The regression yields a significant intercept (95% CI = from 96.052 to 102.824), but no significant slope (95% CI = from -0.098 to 0.135).*

**Table 2:** Population size statistics for the stable trajectory with 10% error

|  |  |  |  |
| --- | --- | --- | --- |
| ***Sample Size*** | ***Mean*** | ***S.D.*** | ***C.V. (%)*** |
| 50 | 99.260 | 28.440 | 28.652 |



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 0.046 |  |  |  |  |  |  |  |
| R Square | 0.002 |  |  |  |  |  |  |  |
| Adj R Square | -0.018 |  |  |  |  |  |  |  |
| Standard Error | 6.106 |  |  |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 3.803 | 3.803 | 0.102 |  0.751 |  |  |  |
| Residual | 49 | 1826.707 | 37.280 |  |  |  |  |  |
| Total | 50 | 1830.510 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower* *95%* | *Upper 95.0%* |
| Intercept | 99.438 | 1.685 | 59.010 | 0.000 |  96.052 | 102.824 |
| year | 0.019 | 0.058 | 0.319 | 0.751 |  -0.098 | 0.135 |



X1The residuals highlight there is more variability in this time series (with 10% measurement error) than in the time series with no error. In both cases, the residuals are superimposed on the flat (slope = 0) best-fit line.

***1c – 2c) Stable, 50% error:*** *This model has very poor fit (r square = 0.011) because the only pattern is random noise (variation). Otherwise, there is no linear trend in abundance over time.*

*The regression yields a significant intercept (95% CI = from 88.526 to 119.901), but no significant slope (95% CI = from -0.738 to 0.343).*

**Table 3:** Population size statistics for the stable trajectory with 10% error

|  |  |  |  |
| --- | --- | --- | --- |
| ***Sample Size*** | ***Mean*** | ***S.D.*** | ***C.V. (%)*** |
| 50 | 99.900 | 6.112 | 6.118 |

******

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 0.104 |  |  |  |  |  |  |  |
| R Square | 0.011 |  |  |  |  |  |  |  |
| Adj R Square | -0.009 |  |  |  |  |  |  |  |
| Standard Error | 28.285 |  |  |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 431.266 | 431.266 | 0.539 | 0.466 |  |  |  |
| Residual | 49 | 39202.891 | 800.059 |  |  |  |  |  |
| Total | 50 | 39634.157 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower* *95%* | *Upper 95%* |
| Intercept | 104.213 | 7.806 | 13.350 | 0.000 | 88.526 | 119.901 |
| year | -0.198 | 0.269 | -0.734 | 0.466 | -0.738 | 0.343 |

******

The residuals highlight there is more variability in this time series (with 50% measurement error) than in the time series with 10% measurement error and with no error. In all cases, the residuals are superimposed on the flat (slope = 0) best-fit line.

***1d – 2d) 1% Decline per Year, No error:*** *This model has perfect fit (explains 100 % of the variance), and detected a significant slope, whose 95% confidence intervals (CI) (from -1 to -1) do not overlap the point estimate of the slope for a null hypothesis (b = 0). The plots of the line and the residuals underscore the perfect fit of this model without noise (no error).*

**Table 4:** Population size statistics for the stable trajectory with 10% error

|  |  |  |  |
| --- | --- | --- | --- |
| ***Sample Size*** | ***Mean*** | ***S.D.*** | ***C.V. (%)*** |
| 50 | 74.500 | 14.577 | 19.567 |



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 1 |  |  |  |  |  |  |  |
| R Square | 1 |  |  |  |  |  |  |  |
| Adj R Square | 1 |  |  |  |  |  |  |  |
| Standard Error | 6.49E-15 |  |  |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 11050 | 11050 | 2.62E+32 | 0 |  |  |  |
| Residual | 49 | 2.06E-27 | 4.21E-29 |  |  |  |  |  |
| Total | 50 | 11050 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower* *95%* | *Upper 95%* |
| Intercept | 100 | 1.79E-15 | 5.58E+16 | 0 | 100 | 100 |
| year | -1 | 6.17E-17 | -1.6E+16 | 0 | -1 | -1 |



***1e – 2e) 1% Decline per Year, 10% error:*** *This model has very good fit (explains 85.4 % of the variance), and detected a significant slope, whose 95% confidence intervals (CI) (from --1.098 to -0.865) do not overlap the point estimate of the slope for a null hypothesis (b = 0). The plots of the line and the residuals underscore the very good fit of this model with some noise.*

**Table 5:** Population size statistics for the stable trajectory with 10% error

|  |  |  |  |
| --- | --- | --- | --- |
| ***Sample Size*** | ***Mean*** | ***S.D.*** | ***C.V. (%)*** |
| 50 | 74.400 | 15.537 | 20.883 |



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 0.924 |  |  |  |  |  |  |  |
| R Square | 0.854 |  |  |  |  |  |  |  |
| Adj R Square | 0.851 |  |  |  |  |  |  |  |
| Standard Error | 6.106 |  |  |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 10643.803 | 10643.803 | 285.512 | 0.000 |  |  |  |
| Residual | 49 | 1826.707 | 37.280 |  |  |  |  |  |
| Total | 50 | 12470.51 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower* *95%* | *Upper 95%* |
| Intercept | 99.438 | 1.685 | 59.010 | 0.000 | 96.052 | 102.824 |
| year | -0.981 | 0.058 | -16.897 | 0.000 | -1.098 | -0.865 |

***1f – 2f) 1% Decline per Year, 50% error:*** *This model has poor fit (explains 28.8% of the variance), and detected a significant slope, whose 95% confidence intervals (CI) (from -1.738 to -0.657) do not overlap the point estimate of the slope for a null hypothesis (b = 0). The plots of the line and the residuals underscore the poor fit of this model with some noise.*

**Table 5:** Population size statistics for the stable trajectory with 10% error

|  |  |  |  |
| --- | --- | --- | --- |
| ***Sample Size*** | ***Mean*** | ***S.D.*** | ***C.V. (%)*** |
| 50 | 73.760 | 33.312 | 45.163 |



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 0.537 |  |  |  |  |  |  |  |
| R Square | 0.288 |  |  |  |  |  |  |  |
| Adj R Square | 0.273 |  |  |  |  |  |  |  |
| Standard Error | 28.285 |  |  |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 15847.266 | 15847.266 | 19.808 |  0.000 |  |  |  |
| Residual | 49 | 39202.891 | 800.059 |  |  |  |  |  |
| Total | 50 | 55050.157 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower* *95%* | *Upper 95%* |
| Intercept | 104.213 | 7.806 | 13.350 | 0.000 | 88.526 | 119.901 |
| year | -1.198 | 0.269 | -4.451 | 0.000 | -1.738 | -0.657 |



***TAKE HOME LESSONS:***

The ‘true trend’ of these data was a linear decrease in abundance (slope = - 1.00 per year). As we increase the amount of error (from 0% to 10% to 50%), the linear trend captures less of the variability in the data, and the range of possible slopes (described by the 95% C.I.) increases.

Yet, the linear decrease was not significant in all three “stable” scenarios and significant in all three ‘declining” scenarios.

To explore how the measurement error (noise) inhibits our ability to detect the trend, we are going to try to see how many years would be necessary to detect a significant decline in the 3 “declining” scenarios, by adding a sampling year at a time to our time series. We will start with 3 years, since we need at least 3 data points to make a line and to calculate the R-square term, which quantifies the amount of the variance explained by the best-fit line.

NOTE: Even though the stable time series without error can yield an R-squared value of 0 (a perfect fit of the data points of the line)… there really is no valid trends line to speak of. The regression is not significant, because the slope = 0. , even if some stats programs may report a significant p value.

* 3) 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 only use the three declining time series and – starting with the first three samples (years 0, 1, 2), 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 decrease for each time series.

For each step, report the p value and the slope (coefficient +/- 95 % confidence intervals). (+1 point for each of the three declining datasets)

Answer these questions:

* **What was the shortest duration you had to sample to determine a significant**

**decrease? What time series were you analyzing? (+1 point)**

***Table 1:*** *Results of the incremental analysis for the time series without error (starting at year 3 and ending at year 12), 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.*

**NOTE:** The time series with no error yielded significance right away. The R-square = 1.

Even though the software does not provide you with an F ratio and p value, this is caused by the lack of residual variance (the denominator = 0). So, what does this mean? That the regression is highly significant rightaway, from the first try using three data points.

My lesson learned: next time I will create three time series with error (1%, 10%, 50%).

***Table 2:*** *Results of the incremental analysis for the fine-scale time series with 10% error (year 0 to 11), 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 those point estimates.*

*We first detect a significant decline at year 6 (with a time series of 7 years).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample Size (years)** | **p value** | **slope coefficient** | **SE slope** | **lower 95% CI** | **upper 95% CI** |
| 3.000 | 0.249 | 3.500 | 1.443 | -14.839 | 21.839 |
| 4.000 | 0.461 | -1.943 | 2.151 | -11.198 | 7.312 |
| 5.000 | 0.138 | -2.651 | 1.319 | -6.852 | 1.549 |
| 6.000 | 0.073 | -2.250 | 0.931 | -4.834 | 0.335 |
| **7.000** | **0.019** | **-2.318** | **0.683** | **-4.075** | **-0.561** |
| **8.000** | **0.003** | **-2.395** | **0.527** | **-3.686** | **-1.104** |
| **9.000** | **0.057** | **-1.546** | **0.680** | **-3.154** | **0.062** |
| **10.000** | **0.035** | **-1.419** | **0.559** | **-2.709** | **-0.128** |
| **11.000** | **0.059** | **-1.236** | **0.571** | **-2.528** | **0.055** |
| **12.000** | **0.026** | **-1.241** | **0.475** | **-2.300** | **-0.183** |

***Table 3:*** *Results of the incremental analysis for the fine-scale time series with 10% error (year 0 to 19), 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 those point estimates.*

*We first detect a significant decline at year 19 (with a time series of 20 years).*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sample Size (years)** | **p value** | **slope coefficient** | **SE slope** | **lower 95% CI** | **upper 95% CI** |
| 3.000 | 0.067 | 11.000 | 1.155 | -3.672 | 25.672 |
| 4.000 | 0.507 | -9.500 | 11.847 | -60.473 | 41.473 |
| 5.000 | 0.975 | -0.300 | 8.660 | -27.860 | 27.260 |
| 6.000 | 0.808 | -1.486 | 5.711 | -17.341 | 14.369 |
| 7.000 | 0.648 | -1.964 | 4.047 | -12.368 | 8.440 |
| 8.000 | 0.428 | -2.583 | 3.038 | -10.017 | 4.850 |
| 9.000 | 0.269 | -2.833 | 2.358 | -8.408 | 2.741 |
| 10.000 | 0.065 | -4.552 | 2.126 | -9.455 | 0.352 |
| 11.000 | 0.220 | -2.691 | 2.042 | -7.309 | 1.927 |
| 12.000 | 0.585 | -1.091 | 1.934 | -5.399 | 3.217 |
| 13.000 | 0.442 | -1.308 | 1.639 | -4.915 | 2.300 |
| 14.000 | 0.250 | -1.721 | 1.424 | -4.823 | 1.381 |
| 15.000 | 0.241 | -1.521 | 1.238 | -4.197 | 1.154 |
| 16.000 | 0.226 | -1.376 | 1.086 | -3.706 | 0.953 |
| 17.000 | 0.535 | -0.662 | 1.043 | -2.885 | 1.561 |
| 18.000 | 0.197 | -1.359 | 1.010 | -3.501 | 0.783 |
| 19.000 | 0.159 | -1.332 | 0.904 | -3.239 | 0.575 |
| **20.000** | **0.044** | **-1.892** | **0.875** | **-3.731** | **-0.053** |

***Figure 4.*** *Results of the incremental analysis of the time series with different levels of error (noise), showing how the point estimate of the slope approaches the real value (b = -1) in both scenarios, but requires different sample sizes to do so. Notice how the highly-variable estimates from the noisy data (50% error) vary widely and do not converge to the correct answer.*

**

***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 because delays for declining populations imply smaller population sizes and worsening conservation outlook. Thus, detecting a trend early is critical to implement conservation actions in a timely fashion.*

*Infrequent 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 decrease 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 decrease? (+1 point)

*The larger the signal (the bigger the slope), the easier it will be to detect a significant decrease, given any specific level of sampling error. The time series with a slope of -4 / year would yield significant results sooner (require smaller sample size), while the time series with a slope of -0.5 / year 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 correctly detect a trend that exists) is an increasing function of the Effect Size (the signal, ES) and a decreasing function of the test statistic (the bigger the alpha level required, the bigger the power). The power is inversely related to the Variability of the data (the noise, V).*

**

*Regarding the actual rate of population increase, the answer is complicated and depends on what is the actual trend of the population.*

*If the population is declining, but at a smaller rate (say the yearly rate of decrease declines from 1% to 0.01%), then the probability of detecting the declining trend will decline. Power will therefore diminish, as the rate of decrease declines.*

*If the population is not changing (not increasing or decreasing, and the slope = 0, and the null hypothesis is true), then the power will be 0. The probability of detecting a decline (or an increase) will be based on the alpha level (commonly 0.05). and will represent a type-I error (erroneous rejection of the true null). Therefore, the power would decline.*

*If the population is increasing, the larger the rate of increase, the larger the power. Power will therefore increase as the rate of increase grows.*