Linear regression model excel

Multi linear regression model in excel. Simple linear regression model excel. Log linear regression model example in excel. Multiple linear regression model excel. How to create a multiple linear regression model in excel. How to build a linear regression model in excel. Nonlinear regression model in excel. Log linear regression model excel.

This January 2009 help sheet provides information on multiple regression using data analysis add-in. Interpret statistical regression. Interpreting the Anova table (often this is jumped.) Interpret the table of regression coefficients. Trust intervals for slope parameters. Test for the statistical meaning of the test hypothesis coefficients on a slope parameter. Test the general meaning of the regraders. Preparing the values of the regraders. Excel limitations. There is little more to know beyond regression with an explanatory variable. The main addition is the F-Test for overall fit. Multiple regressions Using data Analysis Add-In This requires the data analysis addate: see Excel 2007: Access and activating the data analysis addictate The data used are in carsdata.xls then we create a new variable in cells C2: C6, Home size cubbed as regressor. Then in cell C1 give the Cubed Hh Size voice. (It turns out that for data if squared HH Size has a coefficient of exactly 0.0 the cube is used.) The calculation sheet cells A1: C6 must look like: we have a regression with an interception and regressors HH Size e Cubed HH Size The population regression model is: Y = Î²1 + Î²2 x2 + Î²3 X3 + U It is assumed that the u error is independent with constant variation (Homoskedastic) - see Excel Limitations at the bottom. We would like to estimate the regression line: y = b1 + b2 x2 + b3 x3 we do it using the analysis of add-in data and regression. The only modification of the regression is to include more than one column in the X input range. Note, however, that the rickers must be in contiguous columns (here columns B and C.) If this is not the case In the original data, then the columns must be copied to obtain regressors in contiguous columns. Hit ok We get the regression output has three components: Table regression statistics table of table regression coefficients. Interpret statistical table report This is the following exit. Of great interest is R Square. Multiple explanation R 0.895828 R = square root R2 R square 0.802508 R2 adjustment R Square 0.605016 R2 adjusted Used if more than one X Variable Standard Error 0.444401 This is the estimation of the standard deviation sample of the u observations 5 Number of observations used In regression (n) the aforementioned provides general measures of good-of-fit: R2 = 0.8025 Correlation between YE Y-HAT is 0.8958 (when the square 0.8025) R2 adjusted = R2 - (1-R2 * K- 1) / (NK = .8025 - .1975 * 2/2 = 0.6050 The standard error here refers to the estimated standard deviation of the error term u. It is sometimes called the standard regression error. Ie equal sqrt (SSE / (NK.) It is not to be confused with the standard Y-error (from descriptive statistics) or with standard errors of the of regression below. R2 = 0.8025 means that 80.25% of the Yi Yi variation Ybar (its average) is explained by the X2i and X3i regressors. Â, interpreting the Anova table a table is supplied. This is often jumped. Â, DF SS MS F Meaning F Regression 2 1,6050 0.8025 4,0635 0.1975 Residue 2 0.3950 0.1975 Total 4 2.0 The Anova table (variance analysis) divides the sum of the squares into its components. Total amounts of squares = residual sum (or error) of square + regression (or explained) sum of the squares. So Âž £ i (Yi - Ybar) 2 = £ I (Yi - YHATI) 2 + Âž £ I (YHATI - YBAR) 2 Where Yhati is the value of YiÂ ¢ provided by the regression line and Ybar It is the average of the sample. For example: R2 = 1 - SS Residual / Total SS (general formula for R2) Â, Â, Â, Â, Â ¢ = 1 - 0.3950 / 1.6050 Â, Â, Â, Â, Â, Â, Â, Â, Â Â, Â, Â, Â, Â, Â, Â, Â, Â, Â, ã, â â â Â Â, Â, = 0,8025 Â, is, (equal r2 proposed in the regression table statistical). The column labeled F gives the total F-Test test of H0: Î²2 = 0 and Î²3 = 0 against ha: at least one of Î²2 and Î²3 not equal to zero. Apart from: Excel calculates F this as: f = [regression ss / (k-1)] / [residual ss / (n-k)] = [1.6050 / 2] / [.39498 / 2] = 4.0635. The column labeled meaning F has the associated P value. From 0.1975> 0.05, we do not refuse H0 to significance level 0.05. Note: Meaning F In general = Finv (F, K-1, N-K) Â ¢ where K is the number of rickers including HTE interception. Here Finv (4.0635,2,2) = 0.1975. Interpreting the regression of regression coefficients The production of most interests is the following table of coefficients and associated production: Â, coefficient coefficient St. Error T Stat P-value P-value Less 95% Upper 95% intercepted 0.89655 0.76440 1.1729 0.3616 -2.3924 4.1855 HH size 0.33647 0.42270 0.33647 0.42270 0.7960 0.485 -1.4823 2.1552 Cubed HH Size 0.00209 0.01311 0.1594 0.01311 0.1594 0.8880 -0.0543 0.0585 Let Î²J Dende the Rispressor JTH population coefficient (interception, HH size and Cubato HH format). Then "coefficient" column gives estimates of the minimum squares of Î²j. "Standard Error" column provides standard errors (I.E.The estimated standard deviation) of the minimum squares BJ estimates of Î²JJ. "T Stat" column offers statistic T calculated for H0: Î²J = 0 v ° HA: Î²J Â ¢ 0. 0. This is the coefficient divided by standard error. Â compared to with (nk) degrees of freedom where here n = 5 ek = 3. "p-value" column gives the value p for the test of h0: Î²J = 0 against ha: Î²J Â ¢ Â ¢ 0 .. This is equivalent to PR {| T | > T-Stat} Where t is a random variable distributed by T with the N-K grades of freedom and T-Stat is the calculated value of the statistic t indicated in the previous column. Note that this value p is for a front / back test. For a unilateral test divides this value P per 2 (also controlling the sign of the T-Stat). "Lower 95%" columns and "95% superior" values define a 95% confidence interval for Î²JJ. A simple synthesis above is that the line is mounted at a y = 0.8966 + 0.3365 * x + 0.0021 * Z A confidence intervals for inclination coefficients 95% confidence interval for angular coefficient is Â©Â22 Â©Â22 ocita excel (-1.4823, 2.1552). excel calculates this as b2 ± t_.025(3) × if(b2) = 0.33647 ± TINV(0.05, 2) × 0.42270 = 0.33647 ± 4.303 × 0.42270 = 0.33647 ± 1.8189 = (-1.4823, 2.1552). other trust intervals can be obtained. For example, to find 99:% confidence intervals in the regression dialog box (in addition of data analysis,) check the trust level box and set the level to 99.% test hypothesis of zero slope coefficient "(test of hypothetical") the hh size coefficient estimated standard error of 0.4227, 0.7960 t-statistic and 0.5095 p-value. It is therefore statistically insignificant at the meaning level α = .05 as p > 0.05. the hh size cubed coefficient estimated standard error of 0.0131, 0.1594 t-statistic and 0.8880 p-value. It is therefore statistically insignificant at the meaning level α = .05 as p > 0.05. there are 5 observations and 3 regressors (intercept and x) so we oiamo t(5-3)=t(2). For example, for hh size p = TDIST(0.796,2,2) = 0.5095. test hypothesis on a regression parameter here we test if hh size has coefficient β2 = 1.0. example: h0: β2 = 1.0 con ha: β2 ∞ 1.0 at the meaning level α = 05 then t = (b2 - h0 value of β2) / (standard error of b2) = (0.33647 - 1.0) / 0.42270 = -1.569. using the p-value p-value approach = TDIST(1.569, 2, 2) = 0.257. [here n=5 and k=3 so n-k=2]. do not reject the hypothesis at .05 level since the value p is > 0.05. using the critical approach of the value we calculated t = -1.569 the critical value is t_.025(2) = TINV(0.05,2) = 4.303. [here n=5 and k=3 so n-k=2]. therefore do not refuse the hypothesis null at .05 level from t = |-1.569| < 4.303. integral test of the meaning of the parameters of the regressions we test h0: β2 = 0 and β3 = 0 versus has: at least one of the β2 and β3 is not zero. from the table anova the statistics F-test is 4.0635 with p value of 0.1975. Since the p value is not less than 0.05 we do not refuse the hypothesis anything that the regression parameters are zero at the meaning level 0.05. Concluding that the parameters are statistically insignificant at the meaning level 0.05. Note: meaning f in general = FINV(F, k-1, n-k) where k is the number of regreditors including hte interception. here FINV(4.0635.2,2) = 0.1975. value of y regressors consider the case in which x = 4 in which case cubed hh size = x^3 = 4^3 = 64. yhat = b1 + b2 x2 + b3 x3 = 0.88966 + 0.3365×4 + 0.0021×64 = 2.37006 excel limitations excel limits limit the number of regressors (only up to 16 regressors.)? excel requires that all regressor variables be in adjacent columns. you may need to move columns to ensure this. For example. if the regressors are in columns b and d it is necessary to copy at least one of the columns b and d so that they are adjacent to each other. standard excel errors and statistics and p values are basedassumption that the error is independent with a constant changeExcel does not provide alternances, such as heteroskedastic-robust or standard autocorrelations-robust and t-statistic and p-values. You need more specialized software such as STATA, EVIEWS, SAS, LIMDEP, PC-TSP, ... For more information on how to use Excel, please go to We use Excel to adapt the following straight line model to data in example 5.4.1.\[y = \beta_0 + \beta_1 x onumber\] Enter data in a spreadsheet as shown in Figure 5.6.1 . Depending on your needs, there are many ways you can use Excel to complete a linear regression analysis. We will consider three approaches here. If all you need are values for the slope, \(\beta_1\), and the interception y, \(\beta_0\), you can use the following functions: = intercept (known_y's, Know_x's) = slope (known_y's, Know_x's) where Know_y's cell is the range of cells that contain the signals (y), and known_ A7) Excel returns the exact calculation for the slope (120.705 714 3). In order to obtain slope and interception, along with additional statistical details, you can use data analysis tools in the Data Analysis ToolPak. ToolPak is not a standard part of Excel instillation. To see if you have access to the Analysis ToolPak on your computer, select Tools from the menu bar and search for the Data Analysis option.... If you don't see data analysis..., select Add-ins... from the Tools menu. Check the box for the ToolPak Analysis and click OK to install them. Select Data Analysis... from the Tools menu, which opens the Data Analysis window. Scroll through the window, select Regression from the available options and press OK. Place the cursor in the box for the Input Y range and then click and drag the B1 cells: B7. Place the cursor in the box for the Input X range and click and drag the A1 cells: A7. Since the A1 and B1 cells contain labels, check the label box. Including labels is a good idea. Excel's summary output uses the x-axis label to identify the slope. Select the radio button for the output range and click any blank cell; this is where Excel will put the results. Click OK generates the information displayed in Figure 5.6.2 . Figure 5.6.2 : Exit from the Excel Regression command in the Analysis ToolPak. See the text for a discussion on how to interpret the information in these tables. There are three parts to Excel summary of a regression analysis. At the top of figure 5.6.2 is a table of regression statistics. The standard error is the standard deviation on the regression, sr. Also of interest is thefor Multiple R, which is the correlation coefficient of the model, r, a term with which you can already be familiar. The correlation coefficient is a measure of the measure in which the regression model explains the variation in y. The r values range from -1 to +1. TheThe correlation coefficient is Â,Â ± 1, the better the model explains the data. A correlation coefficient of 0 means that there is no relationship between X and Y. In the development of calculations for linear regression, we have not considered the correlation coefficient. There is a reason for this. For most rectilinear calibration curves the correlation coefficient is very close to +1, typically 0.99 or higher. It tends, however, to place too much confidence in the meaning of the correlation coefficient, and to assume that a R greater than 0.99 means that the linear regression model is appropriate. Figure 5.6.3 provides a useful counterexample. Although the regression line has a correlation coefficient of 0.993, the data is clearly curved. The lesson to take home here is simple: don't fall in love with the correlation coefficient! Figure 5.6.3: Example of application of a straight line (in red) to curvilinear data (in blue). The second table of figure 5.6.2 is titled Anova, which is for analysis of the variance. We will give a look more closely to Anova in Chapter 14. For now it is enough to understand that this part of the Excel summary provides information on the fact that the linear regression model explains a significant part of the variation in the values of Y. The value for F is the result of a test F of the following nulle and alternative hypotheses. H0: The regression model does not explain the variation of y ha: the regression model explains the variation of y the value in the column for significance f is the probability of maintaining the hypothesis nothing. In this example, the probability is (2.5 times 10 ^ {- 6}%, which is a strong evidence to accept the regression model. As in the case of the correlation coefficient, a low value for the probability is a probable result for any calibration curve, even when the model is inadequate. The probability of maintaining the hypothesis nothing for figure 5.6.3, for example, is (9.0 times 10 ^ {- 7}%). See Chapter 4.6 For a review of the test F. The third table of figure 5.6.2 provides a synthesis of the model itself. The values for the coefficients of the model Â «The slope, (Beta_1), and the intercept y, (beta_0) Â« are identified as an intercept and with the label for data of the data X axis, which in this example is CSTD. The standard deviations of the coefficients, (s_ {b_0}) and (s_ {b_1}), are in the standard error column. The T Stat column and the p column value are for the following T tests. slope: (h_0 text {:} beta_1 = 0 quad h_a text {:} beta_1 eq 0) y-intercept: (h_0 text {:} beta_0 = 0 quad h_a text { :} beta_0 EQ 0) The results of these T-tests provide convincing tests that the slope is not zero, but there is no evidence that the intercepted y differ significantly from scratch. The 95% confidence intervals are also shown for the slope and the intercept y 95% and 95% higher. See Chapter 4.6 for a test review t. A third approachA regression analysis consists in programming a spreadsheet to manage using the integrated Excel formula for a Sommation = SUM (first cell: last cell) and its ability to analyze mathematical equations. The resulting calculation sheet is shown in Figure 5.6.4. Figure 5.6.4: Computing sheet showing the formulas to calculate the slope and intercept Y for the data of example 5.4.1. Shaded cells contain formulas you need to enter. Insert the formulas in cells from C3 to C7 and in the cells from D3 to D7. Then enter the formulas for cells from A9 to D9. Finally, insert the formulas in cells F2 and F3. When entering a formula, Excel replaces it with the resulting calculation. The values of these cells must comply with the results of example 5.4.1. It is possible to simplify the insertion of the formulas by copying and pasting. For example, enter the formula in cell C2. Select Edit: Copy, click and drag the cursor to cells from C3 to C7, and select Edit: Paste. Excel automatically updates the reference of the cells. You can use Excel to examine the data and the regression line. Start by tracking data. Organize your data in two columns, placing the X values in the leftmost column. Click and drag data and select Graphs from the tape. Select Scatter, choosing the option without lines connecting points. To add a regression line to the chart, click on the graphics of the chart and select Graph: Add trendline ... from the main men. Choose the linear model and click OK to add the line to the graph. By default, Excel displays the regression line from the first to the last point. Figure 5.6.5 shows the result for figure data 5.6.1. Figure 5.6.5: Excellent dispersion graphic example showing data and a regression line. Excel will also create a graph of residual regression model errors. To create the chart, build the regression model using the Toolpak analysis, as described above. Clicking on the option for residual graphics is created the graph shown in Figure 5.6.6. Figure 5.6.6: Example of Excel graph of residual errors of a regression model. The greater limit of Excel for a regression analysis is that it does not provide a function to calculate uncertainty when X values are expected. As for this chapter, Excel cannot calculate the uncertainty for the concentration of the analyte, CA, given the signal for a sample, SSAMP. Another limitation is that Excel does not have an integrated function for a weighted linear regression. However, it is possible to program a spreadsheet to manage these calculations. Exercise 5.6.1 Use Excel to complete the regression analysis in the operation 5.4.1. Answer starts by entering the data in an Excel spreadsheet, following the format shown in the 5.6.1. Because Excel data analysis tools provide most of the information we need, we will use it here. The resulting output, shown below, provides the slope and intercept y, together with the respective 95% confidence intervals. Excel does not provide a function to calculate uncertainty in The analyte concentration, AC, given the signal for a sample, ssamp. You have to complete these calculations by hand. With a ssamp of 0.114, we find that CA is \ [c_a = \ frac {s_ {samp} – b_0} {b_1} = \ frac {0.114 – 0.0014} {29.59 \ text {m} ^ {- 1}}}} = 3.80 \ times 10 ^ {- 3} \ text {m} Oncer \] in CA is \ [s_ {c_a} = \ frac {1.996 \ time 10 ^ {- 3}} {29.59} \ SQRT {\ frac {1} {3} + \ frac {1} {6} + \ frac { (0.114 – 0.1183) ^ 2} { (29.59) ^ 2 \ times 4.408 \ times 10 ^ {- 5}) }} = 4.772 \ Times 10 ^ {- 5} Oncer \] and the 95% confidence interval is \ [\ mu = c_a \ pm ts_ {c_a} = 3.80 \ times 10 ^ {- 3} \ pm \ {2.78 \ times (4.772 \ Times 10 ^ {- 5}) \} Oncer \] \ [\ mu = 3.80 \ times 10 ^ {- 3} 10 ^ {- 3} \ text {m} \ pm 0.13 \ times 10 ^ {- 3} \ text {m} Oncer \ ] Page 2 at a single point we determine the value of KA by measuring the signal for a single standard containing a known concentration of analyte. Using this value of KA and our sample signal, we then calculate the analyte concentration in our sample (see example 5.3.1). With only a single determination of KA, quantitative analysis using external single-point standardization is simple. Multiple point standardization presents a more difficult problem. Consider the data in Table 5.4.1 for external multi-point standardization. What is our best estimate of the relationship between SSTD and CSTD? An attempt is made to treat these data as five separate single point standardizations, determining KA for each standard and reporting the mean value for the five tests. Despite the simplicity, this is not an appropriate way to deal with a multiple-point standardization. Table 5.4.1: Data for a hypothetical multi-point external standardization (c_ {std} \) (arbitrary units) \ (s_ {std} \) (arbitrary units) \ (k_a = s_ {std} / c_ {std} \) 0.000 0.00 â ¬ "0.100 12.36 123.6 0.300 12.36 123.6 0.300 24.83 124.2 0.300 35.83 124.2 0.300 35.91 119.7 0.400 48.79 122.0 0.500 48.79 122.0 0.500 60.42 122.8 Mean KA = 122.5 So why is it inappropriate to calculate an average value for KA using the data in Table 5.4.1? In a single-point standardization, we assume that the reagent blank (the first row in Table 5.4.1) corrects for all constant sources of determined error. If this is not the case, the value of KA from a single point standardization has a constant determining error. Table 5.4.2 shows how an incorrect constant error affects our determination of KA. The first three columns show the analyte concentration in a set of standards, CSTD, the signal without any source of constant error, SSTD and the actual value of KA for five standards. As we expect, the value of KA is the same for each standard. In the fourth column we'll add a fixed constant error +0.50 to signals (SSTD) and. The last column contains the corresponding apparent ka values. Note that we obtain a different value of KA for each standard and that every apparent ka is greater than true value. Table 5.4.2: Effect of a constant error determined on the value of (k_a) from a a Standardization (C_ {STD}) (S_ {STD}) (without constant error) (K_A = S_ {STD} / C_ {STD}) (actual) ((S_ {STD}) _E) (with constant error) 1.10 ka medium (true) = 1.00 ka medium (apparent) = 1.23 How to find the best estimate for the relationship between the signal and the concentration of the analyte in a standardization at multiple points? Figure 5.4.1 shows table data 5.4.1 Tracked as normal calibration curve. Although the data will certainly fall along a straight line, the actual calibration curve is not intuitively obvious. The process to determine the best equation for the calibration curve is called linear regression. Figure 5.4.1: Normal data of the calibration curve for hypothetical external standardization to multiple points referred to in Table 5.4.1. When a calibration curve is a straight line, we represent it using the following mathematical equation [y = beta_0 + beta_1 x label {5.1}], where y is the symbol of the analyte, sstd, ex It is the concentration of the analyte, CSTD. The constants (beta_0) is, respectively, the calibration curve provided for the intercept y and its expected slope. Due to the uncertainty in our measurements, the best we can do is estimate the values of (beta_0) and (beta_1), which we represent as B0 and B1. The objective of a linear regression analysis is to determine the best estimates for B0 and B1. How to do this depends on the uncertainty of our measures. The most common method to complete the linear regression for the equation ref {5.1} includes three hypotheses: that the difference between our experimental data and the regression line calculated is the result of undetermined errors that affect the standard y Standard contributes in equal measure to our estimate of the slope and of the intercept y. For this reason the result is considered a non-weighted linear regression. The second hypothesis is generally true due to the central limit theorem, which we considered in chapter 4. The validity of the two remaining hypotheses is less obvious and it is necessary to evaluate them before accepting the results of a linear regression. In particular the first hypothesis is always suspected because there is certainly some permanent error in the size of X. When preparing a calibration curve, however, it is not rare to find that the uncertainty in the signal, SSTD, is significantly higher than the uncertainty in the concentration of the analyte, CSTD. In such circumstances, the first hypothesis is generally reasonable. To understand the logic of a linear regression, consider the example shown in Figure 5.4.2, which shows three data points and two possible which would reasonably explain the data. How do we decide how well these straight lines fit the data, and how do we determine the best straight line? Figure 5.4.2: Illustration showing three data points and two possible straight lines that could explain the data. The purpose of a linear regression is to find the mathematical model, in this case a straight line, that best explains the data. Let's focus on the solid line in Figure 5.4.2 . The equation for this line is \[\hat{y} = b_0 + b_1 x \label{5.2}\] where b0 and b1 are estimates for the y-intercept and the slope, and \ (\hat{y}\) is the expected value of y for any value of x. Since we assume that all uncertainties are the result of undetermined errors in y, the difference between y and \ (\hat{y}\) for each value of x is the residual error, r, in our mathematical model. \[r_i = (y_i â \hat{y}_i) number\] Figure 5.4.3 shows the residual errors for the three data points. The smaller the total residual error, R, which we define as \[R = \sum_{i = 1}^{n} (y_i â \hat{y}_i) ^2 \label{5.3}\], the better the correspondence between the straight line and the data. In a linear regression analysis, we look for the values of b0 and b1 that give the least total residual error. The reason for squaring individual residual errors is to prevent a positive residual error from canceling out a negative residual error. You've already seen this in the equations for the standard deviations of the sample and the population. You can also see from this equation why a linear regression is sometimes called the least squares method. Figure 5.4.3: Illustration showing the evaluation of a linear regression where all uncertainty is assumed to be the result of indeterminate errors in y. The blue dots, y , are the original data and the red dots, yi , are the values predicted by the regression equation, \ (\hat{y} = b_0 + b_1 x\).The smaller the total residual error (Equation \ref{5.3}), the better the correspondence of the straight line with data. Although we will not formally develop mathematical equations for linear regression analysis, derivatives can be found in many standard statistical texts [See, for example, Draper, N. R.; Smith, H. Applied Regression Analysis, 3rd ed.; Wiley: New York, 1998]. The resulting equation for the slope, b1, is \[b_1 = \frac {n \sum_{i = 1}^{n} x_i y_i â \sum_{i = 1}^{n} x_i \sum_{i = 1}^{n} y_i} {n \sum_{i = 1}} {n} x_i = 1} 2 â \left (\sum_{i = 1}^{n} x_i \right) ^2} \label{5.4}\] and the equation of the y, b0, is \[b_0 = \frac {\sum_{i = 1}^{n} y_i â b_1 \sum_{i = 1}^{n} x_i} {n} \label{5.5}\] Although the equations \ref{5.4} and \ref{5.5} look formidable, you need to evaluate only the following four summations \[\sum_{i = 1}^{n} x_i \quad \sum_{i = 1}^{n} y_i \qquad \sum_{i = 1}^{n} x_i y_i \quad \sum_{i = 1}^{n} x_i^2 number\] to perform a linear regression analysis based on this model. model.Save time and to avoid tedious calculations, learn how to use one of these tools (and see Section 5.6 for details on completing a linear regression analysis using Excel and R.). For illustrative purposes, the required calculations are shown in detail in the following example. Equation \ Ref {5.4} and equation \ ref {5.5} are written in terms of general variables x and y. As you work through this example, remember that x corresponds to CSTD and Y corresponds to SSTD. EXAMPLE 5.4.1 Using data from Table 5.4.1, determine the relationship between SSTD and CSTD using an integrated linear regression. Solution Let's start by setting up a table to help us organize the calculation. \ (x_i \) \ (y_i \) \ (x_i y_i \) 0.000 0.00 0.000 0.000 0.100 12.36 1.236 0.010 0.200 24.83 4.966 0.040 0.300 35.91 10.773 0.090 0.400 48.79 19.516 0.160 0.500 60.42 30.210 0.250 Value addition In each column gives \ [\ sum_ {i = 1} ^ {n} x_i = 1,500 \ quad \ sum_ {i = 1} ^ {n} y_i = 182.31 \ quad \ sum_ {i = 1} ^ {n} x_i y_i = 66,701 \ quad \ sum_ {i = 1} ^ {n} x_i ^ 2 = 0.550 onumber \] Replace these values with equation \ ref {5.4} and equation \ ref {f { 5.5}, we find that the slope and the y-intercept are \ [B_1 = \ frac { 6 \ times 66.701) – (1,500 \ times 182.31) } { (6 \ times 0.550) – (1.500) ^ 2} = 120.706 \ about 120.71 Onumber \] \ [B_0 = \ frac {182.31 – (120,706 \ times 1,500) } {6} = 0.209 \ about 0.21 Oz. \] The relationship between the signal and the analyte, then, is \ [s_ {std} = 120.71 \ times c_ {std} + 0.21 Oz. \] For now we keep two decimal points to match the number of decimal places in the signal. The resulting calibration curve is shown in Figure 5.4.4. Figure 5.4.4: Calibration Curve for Data in Table 5.4.1 and Example 5.4.1. As shown in Figure 5.4.4, because of the indeterminate errors in the signal, the regression line might not pass through the exact center of each data point. The cumulative deviation of our regression line data – that is, the total residual error – is proportional to the uncertainty in the regression. We call this uncertainty the standard deviation on regression, SR, which is equal to \ [s_r = \ sqrt {\ frac {\ sqrt {\ frac {\ sum_ {i = 1} ^ {n} \ left (y_i – \ hat {y} _i \ Right) ^ 2} {n – 2}} {n-2}}}} \ label {5.6} \] Where Yi is the experimental value Ith, and \ (\ hat {y} _i \) is the corresponding value predicted by the regression line in equation \ ref {5.2}. Note that the denominator of the equation \ def {5.6} indicates that our regression analysis has n – 2 degrees of freedom "we lose two degrees of freedom because we use two parameters, slope and intercept Y, to calculate \ (\ cap {y} _i \). Regression standard deviation (Equation \ ref {5.6}) and standard deviation for a sample (Equation 4.1.1)? A more useful representation of uncertainty in our regression is to consider the effect of indeterminate errors on the slope, B1 and the intercept Y, B0, which we express as standard deviations. \ [s_ {b_1} = = }========================= We use these standard deviations to establish confidence intervals for the predicted slope, \ (\beta_1\), and the predicted y-intercept, \ (\beta_0\) \[\beta_1 = b_1 \pm t s_{b_1} \label{5.9}\] \[\beta_0= b_0\pm t s_{b_0} \label{5. Note that Equation \ref{5.9} and Equation \ref{5.10} do not contain a \ (\sqrt{n}) ^{-1}\) factor because the confidence interval is based on a single regression row. Example 5.4.2 Calculate 95% confidence intervals for slope and intercept example 5.4.1 . Solution Let's start by calculating the standard deviation on the regression. To do this we need to calculate the above signals, \ (\hat{y}_i\), using the slope and y-intercept from example 5.4.1 , and the squares of the residual error, \ (y_i - \hat{y}_i) ^2\). Using the last standard as an example, we find that the predicted signal is \[\hat{y}_6 = b_0 + b_1 x_6 = 0.209 + (120.706 \times 0.500) = 60.562 onumber\] and that the remaining error square is \[(y_i - \hat{y}_i) ^2 = (60.appro – 60.02x \ (x_i\) \ (y_i\) \ (\) \ (\hat{y}_i\) \ (\) \\\\ \ (\left (y_i – \hat{y}_i \right) ^2\) 0.000 0.00 0.209 0.0437 0.100 12.36 12.280 0.0064 0.200 24.350 0.2304 0.300 Adding together the data in the last column gives the numerator of Equation \ref{5.6} as 0.6512; hence, the standard deviation on the regression is \[s_r = \sqrt{\frac {0.6512} {6 – 2}} = 0.4035 onumber\] Next we calculate the standard deviations for p trend and y intercept using Equation \ref{5.7} and Equation \ref{5.8}. The values of the summation terms are from Example 5.4.1 . \[s_{b_1} = \sqrt{\frac {6 \times (0.4035) ^2} { (6 \time 0.550) – (1.500) ^2}} = 0,65 onumber\] \[s_{b_0} = \sqrt{\frac { (0.4035) ^2 \time 0.550} { (6 \time 0.550) – (1.500) ^2}} = 0.292 onumber\] Finally, the 95% confidence intervals (\ (\alpha = 0.05\), 4 degrees of freedom) for slope and y-intercept are \[\beta_1 = b_1 \pm ts_{b_1} = 120.706 \pm (2.78 \times 0.965) = 120.7 \pm 2.7 onumber\] \[\beta_0 = b_0 \pm ts_{b_0} = 0.209 \pm (2.78 \time 0.292) = 0.2 \pm 0.80 onumber\] where t (0.05, 4) from Appendix 4 is 2.78. The standard deviation on the regression, sr, suggests that the signal, Sstd, is accurate to one decimal point. For this reason, the slope and the intercept are reported to a single decimal point. To minimize uncertainty in a calibrationSlope and Y-Intercept, we evenly space our standards on a wide range of analytical concentrations. An in-depth examination of the equation ref {5.7} and the equation ref {5.8} help us understand why it is true. The denominators of both equations include the term (sm_ {i = 1} ^ {n} (x_i - overline {x} _i) ^ 2). Biggest is the value of this term, which we carry out by increasing the range of x around its average value "smaller the standard deviations in slope and interception y. Furthermore, to minimize uncertainty in Y interception, it helps to decrease the value of the term (sm_ {i = 1} {n} x_i) in equation ref {5.8}, which we are including the standards For lower concentrations of the analyte. Once we have our regression equation, it is easy to determine the concentration of analytes in a sample. When we use a normal calibration curve, for example, measure the signal for our sample, SSAMP, and we calculate the concentration of the analyte, CA, using the regression equation. [C_a = frac {s_ {samp} - b_0} {b_1} label {5.11} that is less obvious is how to report a trust interval for ca expressing uncertainty in our analysis. To calculate a trusted interval we need to know the standard deviation in the concentration of the analyte, which is given by the following equation [s_ {c_a} = frac {s_r} {b_1} sqrt {frac {1} {m} + frac {1} {n} + frac {left (soverline {s} - overline {c} _ {std} right) ^ 2}} label {5.12 } Where M is the number of replication we use to establish the average sample signal, SSAMP, N is the number of calibration standards, SSTD is the average signal for calibration standards, and (C_ { STD_1}) E (Overline {C} _ {STD}) are individual concentrations. Knowing the value of (s_ {c_a}), the trust interval for the concentration of the analyte is [mu_ {c_a} = c_a pm t s_ {c_a} Onumber where (mu_ {C_A} is the expected value of CA in the absence of certain errors, and with the value of t is based on the desired trust level en â € "2 degrees of freedom. Equation Ref {5.12} is written in terms of a calibration experiment. A more general form of the equation, written in terms of X and Y, is given here. [s_ {x} = frac {s_r} {b_1} sqrt {frac {1} {m} + frac {1} {n} + frac {left ({y} - overline {y}}} 2} {(b_1)} 2 sum_ {i = 1} ^ {n} left (x_i - overline {x} an in-depth examination of the equation ref {5.12} should convince you that l 'CA uncertainty is smaller when the average sample signal, (overline {s} _ {samp}), is the same as the average signal for standards, (overline {s} _ {STD} - (b) Sharaf, A.; Illman, D. L.; Kowalski, B. R. Chemometrics, Wiley-InterScience: New York, 1986, pp. 126-127; (c) Committee of analytical methods Â ¢ â,¬ Â ¢ â,¬ å "uncectures in concentrations estimated by calibration experiments, Â ¢ â,¬" AMC Technical Brief, March 2006. Example 5.4.3 Three replicated analyzes for a sample containing an unknown concentration of analyte, performance values for ssamp of 29.32, 29.16 and 29.51 (arbitrary units). Using the results from example 5.4.1 and example 5.4.2, determine the concentration of analyte, CA and its 95% confidence interval. Solution The average signal, (overline {s} _ {samp}), is 29.33, which, using the equation ref {5.11} and slope and interception 5.4.1, Due the analyzed "s concentration as [c_a = frac {overline {s} _ {samp} - b_0} {b_1} = frac {29.33 - 0.209} {120.706} = 0.241 to calculate the standard deviation for the Â "¢ s concentration {s {s} _ {STD}} {{}} {} {} {2} (C_ {STD_I} - Overline {C} _ { STD}) ^ 2). The first is only the average signal for calibration standards, which, using data in table 5.4.1, is 30.385. Calculation (} {i = 1} ^ {2} (c_ {std_i} - overline {c} _ {std}) ^ 2 {std}) ^ 2) seems to be formidable, but we can simplify your calculation Recognizing that this sum of -soles is the numerator in a standard deviation equation; Then, [sum_ {i = 1} ^ {n} (c_ {std_i} - overline {c} _ {std}) ^ 2 = (s_ {c_ {std}}) ^ 2 times (n - 1) Oncer, where (s_ {C_ {STD}}) is the standard deviation for the analyte concentration in calibration standards. Using data in Table 5.4.1 We find that (s_ {c_ {STD}}}) is 0.1871 and [sum_ {i = 1 ^ {n} (c_ {std_i} - overline {c } _ {STD}} ^ 2 = (0.1872) ^ 2 times (6 - 1) = 0.175 Oncer, replacement of known values in the equation ref {5.12} = "0.4035 c_a} = } {120.706} sqrt {frac {1} {3} + frac {1} {6} + frac {(29.33 - 30.385) ^ 2} {(120.706) ^ 2 times 0.175}}} = 0.0024 Finally, the 95% confidence interval for 4 degrees of freedom is [mu_ {c_a} = c_a pm ts_ {c_a} = 0.241 pm (2.78 times 0.0024) = 0.241 pm 0.007 oncer Figure 5.4.5 shows the calibration curve with curves that show the 95% confidence interval for approx. Figure 5.4.5: Example of a normal calibration curve with a confidence interval superimposed for the concentration of the analyte. The points in blue are the original data from Table 5.4.1. The black line is the normal calibration curve determined in Example 5.4.1. The red lines show the 95% confidence interval for CA by assuming a single determination of the SSAMP. In a standard addition we determine the concentration of the analyte extrapolating the calibration curve to the X interception. In this case the value of CA is [C_a = x text {-ertercept} = frac {-b_0} Oncer and the standard deviation in CA is \ [s_ {c_a} = \ frac {s_r} {B_1} \ sqrt {\ frac {1} {n} + \ frac {(\ overline {s} _ {std} {B_1})}{\sum_ {i = 1} {n} {n}(C_ {std_i})}of standard additions (including the sample without additional standard), and \(\overline{S}_{std}\) is the average signal for standard n. Since the concentration of analyte by extrapolation is determined, rather than by interpolation, \(s_{C_A}\) for the standard addition method is generally larger than a normal calibration curve. Exercise 5.4.1 Figure 5.4.2 shows a normal calibration curve for the quantitative analysis of Cu2+. Here are the data for the calibration curve. [Cu2+] (M) Absorption 0 0 \(1.55 \times 10^{-3}\) 0.050 \(3.16 \times 10^{-3}\) 0.093 \(4.74 \times 10^{-3}\) 0.143 \(6.34 \times 10^{-3}\) 0.188 \(7.92 \time 10^{-3}\) If three replicated samples give a 0.114 Ssamp, what is the concentration of analytics in the sample and its 95% confidence interval? \\ Adding the values in each column gives \[\sum_{i = 1}^{n} x_i = 2.371 \times 10^{-2} \quad \sum_{i = 1}^{n} y_i = 0.710 \quad \sum_{i = 1}^{n} x_i = 4.110 \time 10^{-3} \quad \sum_{i = 1}^{n When we replace these values in Equation \ref{5.4} and Equation \ref{5.5}, we find that the gradient and interception y are \[b_1=\frac {6 \times (4.110 \37times 10^{-3}) - (2.371 \times 10^{-2}\times 0.710} {6 \times (1.378 \stime To calculate 95% confidence intervals, we must first determine the standard deviation on regression. The following table helps us organize the calculation. \5(x_i)\(y_i)\(\hat{y}_i)\(y_i_}^{_i} } 0,000 0,000 0.0015 \(2.250 \time 10^{-6}\)\1.5(1.55time 10^{-3}\) 0.050 0.0473 \ Adding together the data in the last column gives the Equation \5 The standard deviation on regression, therefore, is \[s_r = \sqrt{\frac {1.596 \times{6 â € "2}} = 1.997 Times 10 ^ {- 3} Number subsequently, we must calculate the standard deviations for the slope and the intercept Y using the equation ref {5.7} and the equation ™ Equation ref {5.8}. [s_ {b_1} = sqrt {frac {6 times (1.997 2} {6 time (1.378 time 10 {- 4}) â € "(2.371 Time 10 ^ {- 2}) ^ 2}} = 0.3007 in Number ^^^^^^^^^ {-4})} {6 times (1.378 times 10 ^ {- 4}) â € "(2.371 times 10 ^ {- 2}) ^ 2} } = 1.441 Times 10 ^ {- 3} Number and USALS To calculate the 95% confidence intervals for the slope and the intercept y [beta_1 = b_1 PM TS_ {B_1} = 29.57 PM (2.78 times 0.3007) = 29.57 text {m} ^ {- 1} 0.84 text {m} ^ {- 1} Number] [beta_0 = b_0 pm ts_ {b_0} = 0.0015 pm (2.78 TIME 1.441 TIME 10 ^ {- 3}) = 0.0015 pm 0.0040 Number with an average SSAMP of 0.114, the concentration of the analyte, CA, is [C_A = Frac {s_ {SAMP } â € "b_0} {b_1} = frac {0.114 â € œ 0.0015} {29.57 text {m} ^ {- 1}} = 3.80 Times 10 ^ {- 3} text {m} Number The standard CA deviation is [s_ {c_a} = frac {1.997 times 10 ^ {- 3}} {29.57} sqrt {frac {1} {3} + frac {1} {6} + Frac {(0.114 â € œ1 0.1183) ^ 2} {(29.57) ^ 2 Times (4.408 times 10 ^ {- 5})}} = 4.778 Times 10 ^ {- 5} Number and the 95% confidence interval is [MU = C_A PM T S_ {C_a} = 3.80 Times 10 ^ {- 3} pm {2.78 Times (4.778 Times 10 ^ {- 5}) } Text {m} PM 0.13 Times 10 ^ {- 3} text {m} Number You should never accept the result of a linear regression analysis without evaluating the validity of the model. Perhaps the simplest way to evaluate a regression analysis is to examine residual errors. As we have seen earlier, the residual error for a single calibration standard, RI, is [R_I = (y_i â € "HAT {Y} _I) Number if the regression model is valid, Then residual errors should be randomly distributed on an average residual error of zero, without any apparent trend towards minor or greater residue errors (figure 5.4.6 a). Trends such as those of Figure 5.4.6 B and Figure 5.4.6 C show that at least one of the model hypotheses is incorrect. For example, a trend towards larger residual errors at higher concentrations, figure 5.4.6 B, suggests that undetermined errors that influence the signal are not independent of the concentration of the analyte. Figure 5.4.6 C, residual errors are not random, which suggests that we cannot model data using a linear report. The regression methods for the last two cases are discussed in the following sections. Figure 5.4.6: Graphs of the residual error in the signal, SSTD, depending on the concentration of the analyte, CSTD, for a non-weighted linear regression model. The red line shows a residual error of zero. The distribution of residual errors referred to in letter a) indicates the adequacy of the non-weighted linear regression model. The increase residual errors referred to in point (b) for higher analyte concentrations suggest that a weighted weighted line It is more appropriate for (C), the curved residue model suggests that a linear model is inappropriate; Linear regression with a quadratic model could produce a better fit. Exercise 5.4.2 Use of results from exercise 5.4.1, build a residual texture and explains its meaning. Answer To create a residual texture, we need to calculate the remaining error for each standard. The following table contains relevant information. (x_i) (y_i) (y__}_i) ^ {- 3} 0,050 0.0473 0.0027 (3.16 times 10 ^ {- 3} 0.093 0.0949 Â ¢ â,¬ "0.0019 (4.74 times 10 ^ {- 3}) 0.143 0.1417 0 , 0013 (6.34 times 10 ^ {- 3} 0.188 0.1890 Â ¢ â,¬ "0.0010 (7.92 times 10 ^ {- 3}) 0.236 0.2357 0.0003 The following figure shows a graph of the resulting residual errors . The residual errors appear random, although they alternate in the sign and that they show no significant dependence on the concentration of the analyte. Taken together, these observations suggest that our regression model is appropriate. Our treatment of linear regression at this point it assumes that the undetermined errors that affect Y are independent of the value of X. If this hypothesis is false, as in the case of data in Figure 5.4.6 B, so we must include the variance for each value of Y in ours determination of Y, B0 and slope interception, B1; then [b_0 = frac {sm_ {i = 1} ^ {n} w_i y_i - b_1 sm_ {i = 1} ^ {n} w_i x_i} {n} label {5.13}] [b_1 = frac {n} {i = 1} ^ {n} w_i x_i y_i - sum_ {i = 1} ^ {n} w_i x_i sum_ {i = 1} {n} w_i y_i} {n sum_ {i = 1} ^ {n} w_i x_i ^ 2 - left (sum_ {i = 1} {n} w_i x_i right) ^ 2} label {5.14} where wi is a factor of weighting explaining the variance in Yi [w_i = frac {n (s_ {y_i}) ^ {- 2}} {m_ {i = 1} {n} (s_ {y_i}) ^ {- 2}} label {5.15} and (s_ {y_i}) is the standard deviation for yi. In a weighted linear regression, the contribution of each pair of XY to the regression line is inversely proportional to the precision of Yi; That is, the more precise the value of Y, the greater its contribution to regression. Example 5.4.4 The following are the data for external standardization in which SSTD is the standard deviation for three replicas signal determinations. This is the same data used in Example 5.4.1 with more information on standard deviations in the signal. (C_ {STD}) (arbitrary units) (s_ {STD}) (arbitrary unit) 0.000 0.00 0.02 0.100 12.36 0.02 0.200 24.83 0.07 0.300 35.91 0.13 0.400 48.79 0.32 0.500 60.42 0.33 Determines the calibration curve equation using a weighted linear regression. While working through this example, remember that x corresponds to CSTD and that Y matches SSTD. Solution Let's start A table to help calculate weighting factors. (C_ {STD}) (arbitrary unit) (s_ {STD}) (arbitrary units) (S_ {STD}) ^ {- 2}) (W_i) 0.00 0.00 0.02 2500.00 2.8339 0.100 12.36 0.02 250.00 2.8339 0.200 24.83 0.07 204.08 0.2313 0.2313 35.91 0.13 59,17 0.0671 0,400 48,79 0.22 20,66 0.0234 0.500 60,42 0.33 9,18 0.0104 Together the values of the fourth column are obtained \[\sum_{i = 1}^{n} (s_{y_i}) ^{-2} number\] which we use to calculate individual weights in the last column. To verify your calculations, the sum of individual weights must be equal to the number of calibration samples, no. The sum of entries in the last column is 6.0000, so it's all right. After calculating individual weights, we use a second table to help calculate the four sum terms in the \ref{5.13} and \ref{5.14} equations. < > The calibration equation is \[S_{std} = 122.98 \time C_{std} + 0.2 number\] Figure 5.4.7 shows the calibration curve for the weighted regression and the calibration curve for the unweighted regression in example 5.4 1 . Although the two calibration curves are very similar, there are slight differences in the slope and in the y intercetta. In particular, the intercept y for the linear weighted regression is closer to the expected value of zero. As standard signal deviation, Sstd, is less for lower analytic concentrations, Cstd, a weighted linear regression gives more emphasis to these standards, allowing better estimation of the y intercept. Figure 5.4.7: Comparison of normal calibration curves do not weigh and weigh. See example 5.4.1 for details of unweighted linear regression and example 5.4.4 for details of the weighted linear regression. The equations for calculating the confidence intervals for slope, intercept y and concentration of the analyte when using a linear regression weighted are not so easy to define as for a linear regression not weighted [Bonate, P. J. Anal. Chem. 1993, 65, 1367 [1372]. However, the confidence interval for the concentration of the analyte reaches its optimal value when the signal of the analyte is located near the weighted, yc centerpiece of the calibration curve. \[y_c = \frac {1}\sum_{i = 1}^{n} w_i x_i number\] undetermined errors that affect a calibration curveIn the signal (Y), therefore we must also consider the regression model the undetermined errors that influence the concentration of the analyte in calibration standards (X). The solution for the resulting regression line is computationally more involved than that for the weighted or weighted regression lines. Although we will not consider the details in this textbook, you should be aware that neglecting the presence of indefinite errors in x can affect the results of a linear regression. See, for example, the Committee of Analytical Methods, Â ¢ â,¬ å "Fitting a linear functional relationship with error data both of both variables, Â ¢ â,¬ AMC Technical Brief, March, 2002), as well as of these additional resources of the chapter. A rectilinear regression model, despite its apparent complexity, is the simplest functional relationship between two variables. What do we do if our calibration curve is curvilinear ... ie, if it's a curved line instead of a straight line? An approach is to try to transform data into a straight line. In this way, logarithms, exponentials, mutual, square roots and trigonometric functions were used. Log ground (Y) against X is a typical example. These transformations are not without complications, of which the most obvious is that the data with a uniform variance in Y will not maintain that uniform variance after it becomes. It is worth noting that the term Â ¢ â,¬ Å "LinearÂ ¢ â,¬ does not mean a straight line. A linear function can contain more than one additive term, but each end of the genre has one and a single adjustable multiplicative parameter. The [Y = AX + BX ^ 2 OnBery function is an example of a linear function since the terms X and X2 each include a single multiplicative parameter, A and B, respectively. The function [y = x ^ b OnBerry] is not linear because B is not a multiplicative parameter; Instead, it is a power. This is why you can use linear regression to adapt to a polynomial equation to your data. Sometimes it is possible to transform a non-linear function into a linear function. For example, taking the register of both sides of the non-linear function above provides a linear function. [L log (Y) = b Log (X) Onumber Another approach to the development of a linear regression model is to adapt a polynomial equation to the data, such as (Y = A + BX + CX ^ 2 . You can use the linear regression to calculate the parameters A, B and C, although the equations are different from those for the linear regression of a straight line. If your data cannot be mounted using a single polynomial equation, it may be possible to adapt to the separate polynomial equations to the short segments of the calibration curve. The result is a single continuous calibration curve known as a spline function. For details on regression see (a) Sharaf, M. a.; Illman, D. L.; Kowalski, B. R. Chemometrics, Wiley-IntersCience: New York, 1986; (b) Deming, S. N.; Morgan, S. L. Experimental Design: a chimometric approach, Elsevier: Amsterdam, 1987. Regression models in this thisonly apply to functions that contain a single independent variable, such as a signal that depends on the concentration of the analyte. In the presence of an interferor, however, the signal may depend on the concentrations of both the analyte and the interferor \[S = k_A C_A + k_I CI + S_{reag} number\], where kI is the sensitivity of the interferor and CI is the concentration of the interferor. Multivariate calibration curves are prepared using standards containing known quantities of analyte and interferer and modeled using multivariate regression. See Beebe, K. R.; Kowalski, B. R. Anal. Chem. 1987, 59, 1007A1017A. for more details, and see this Additional Resources chapter for more information on linear regression with errors in both variables, curvilinear regression and multivariate regression. regression.