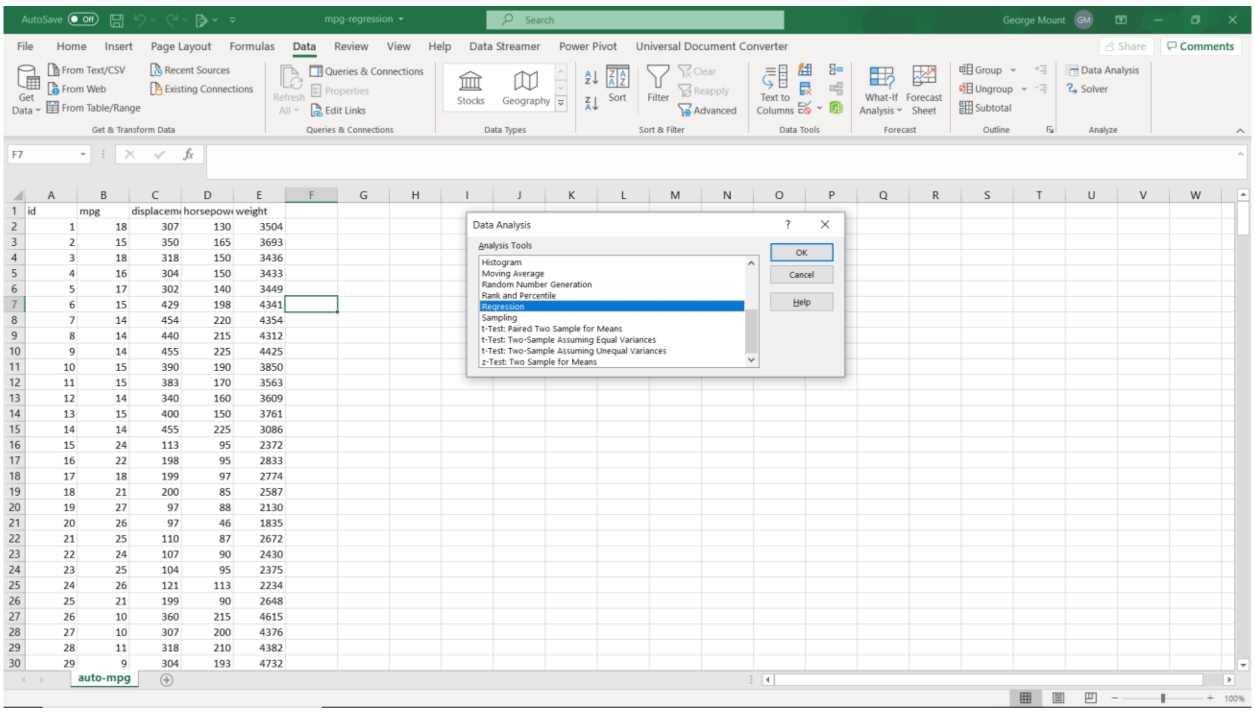
# Regression analysis and predictive models: demo notes

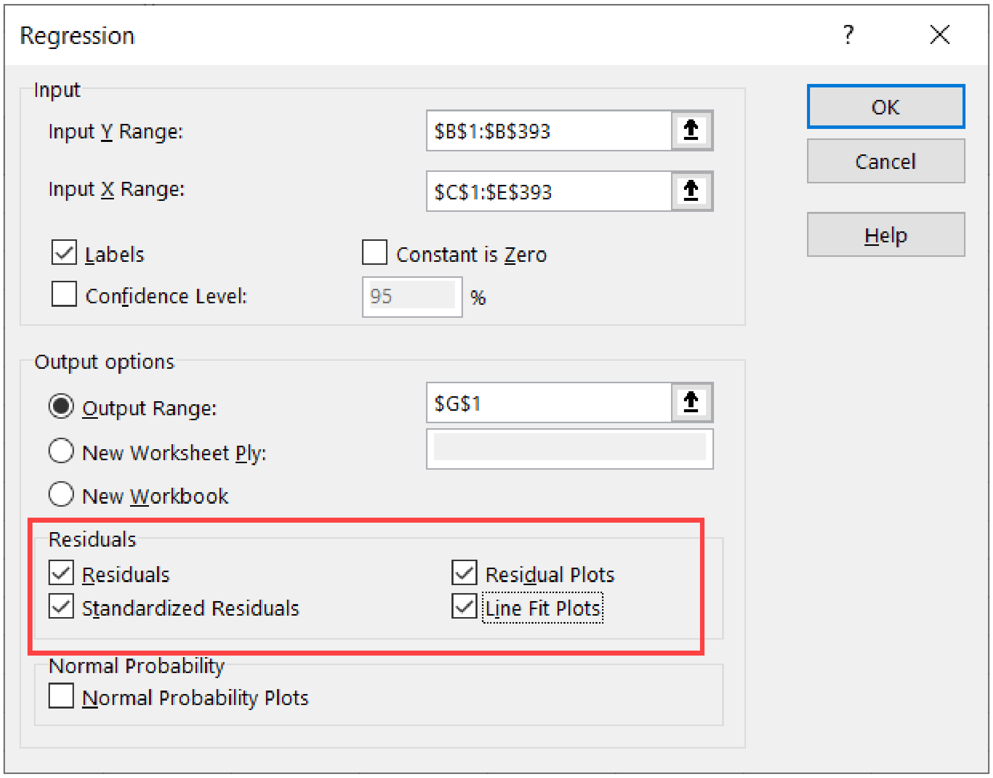
**Multiple linear regression**

Demo: `mpg-regression.xlsx`

1. Head to Data > Data Analysis > Regression.



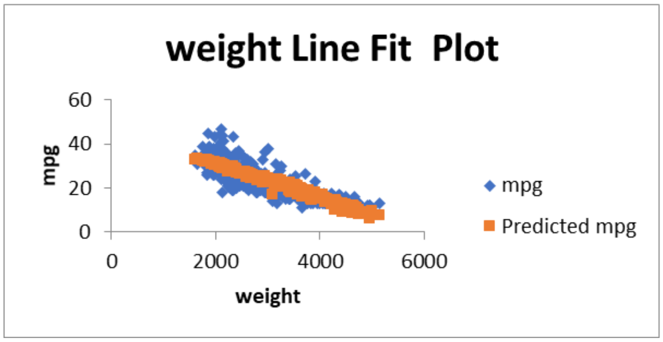

2. The input Y range is our dependent variable, `mpg`. The X range is our three independent variables: `displacement`, `horsepower` and `weight`.
   a. Check all boxes in the Residuals group.

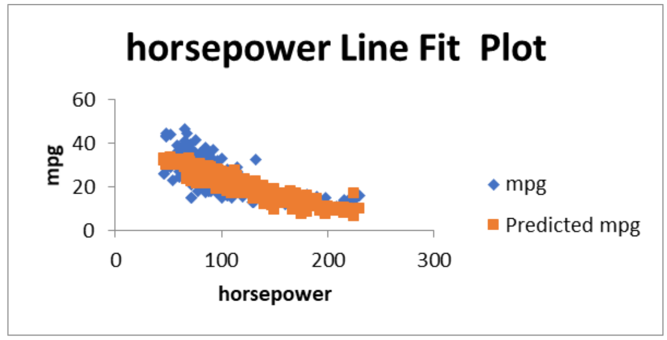



3. We will start our analysis by checking the p-values of our regression, and dropping any non-significant variables.
    a. `displacement` is a non-significant-value, so we will drop it from our next round.

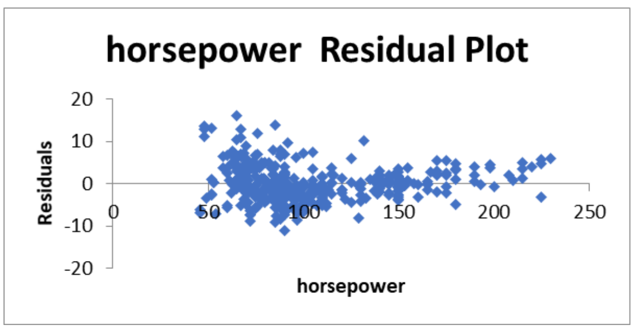
| | Coefficients | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
|---|---|---|---|---|---|---|---|---|
| Intercept | 44.8559357 | 1.195920013 | 37.50747142 | 4.1075E-131 | 42.50464109 | 47.2072303 | 42.50464109 | 47.2072303 |
| displacement | -0.005768819 | 0.00658189 | -0.876468422 | 0.381317748 | -0.018709452 | 0.007171815 | -0.018709452 | 0.007171815 |
| horsepower | -0.041674144 | 0.012813862 | -3.252270314 | 0.001245063 | -0.066867438 | -0.016480849 | -0.066867438 | -0.016480849 |
| weight | -0.005351593 | 0.000712354 | -7.512546915 | 4.03584E-13 | -0.00675215 | -0.003951036 | -0.00675215 | -0.003951036 |

4. Make a copy of the worksheet and delete over the old output. This time only include horsepower and weight in your X range. Leave the rest of the regression settings as-is.
    a. All variables are now significant 🎉. Let's continue in interpreting the results given by the plots we selected to include.
5. First we get two plots of our independent variables with our actual versus predicted Y variable. We can visually see that there is indeed a line fit into each of the scatterplots. This is an assumption of regression.



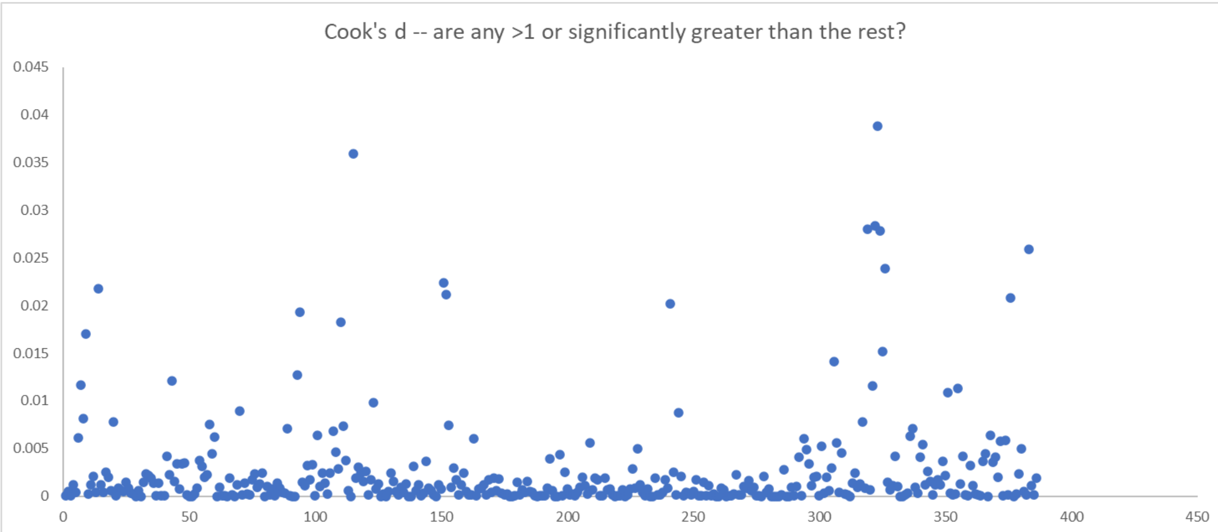




6. We also got two plots of our independent variables against the *residuals*. Remember, these are supposed to look totally *random*, but there is a big clump in the left of each. We may have a problem meeting this assumption.



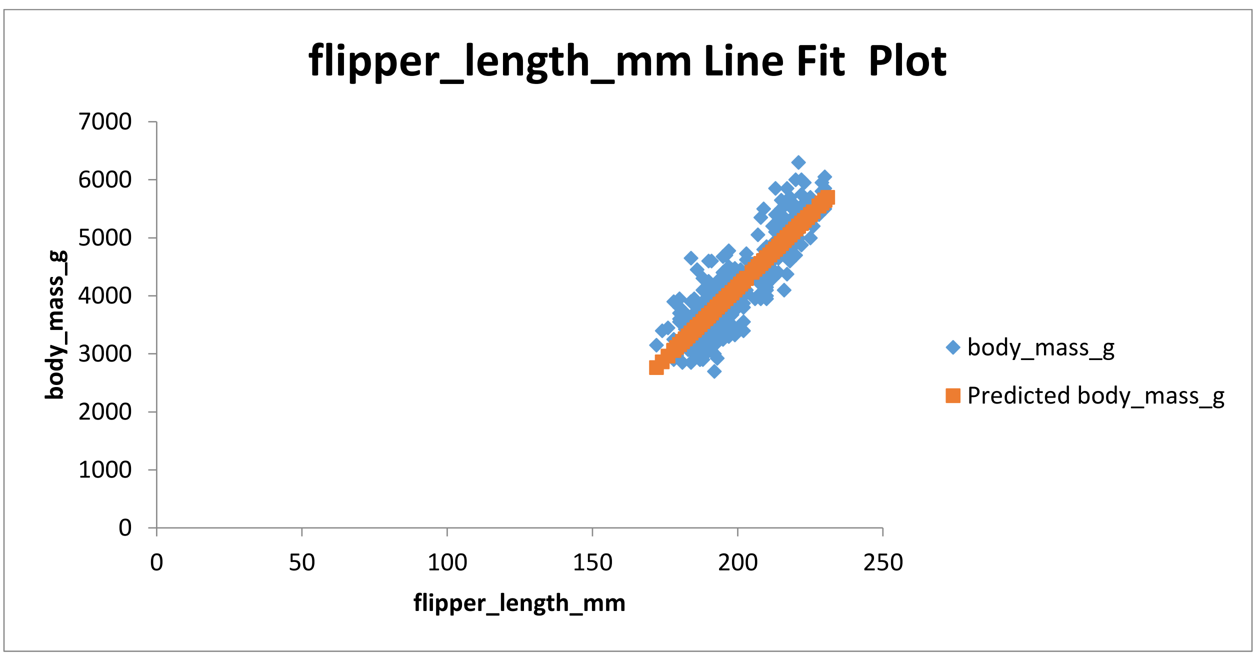


7. Finally, we want to check for influential cases. This is possible to do in Excel but takes some heavy set-up. Un-hide the `influential-cases` worksheet in your workbook.
    a. I have calculated a measure called Cook's D to check for influential cases. Generally a Cook's D greater than 1 signals an influential case, however if there are any much different than the others, those could be considered influential.
        i. In this case, it may be worth identifying two cases as possible influential cases.

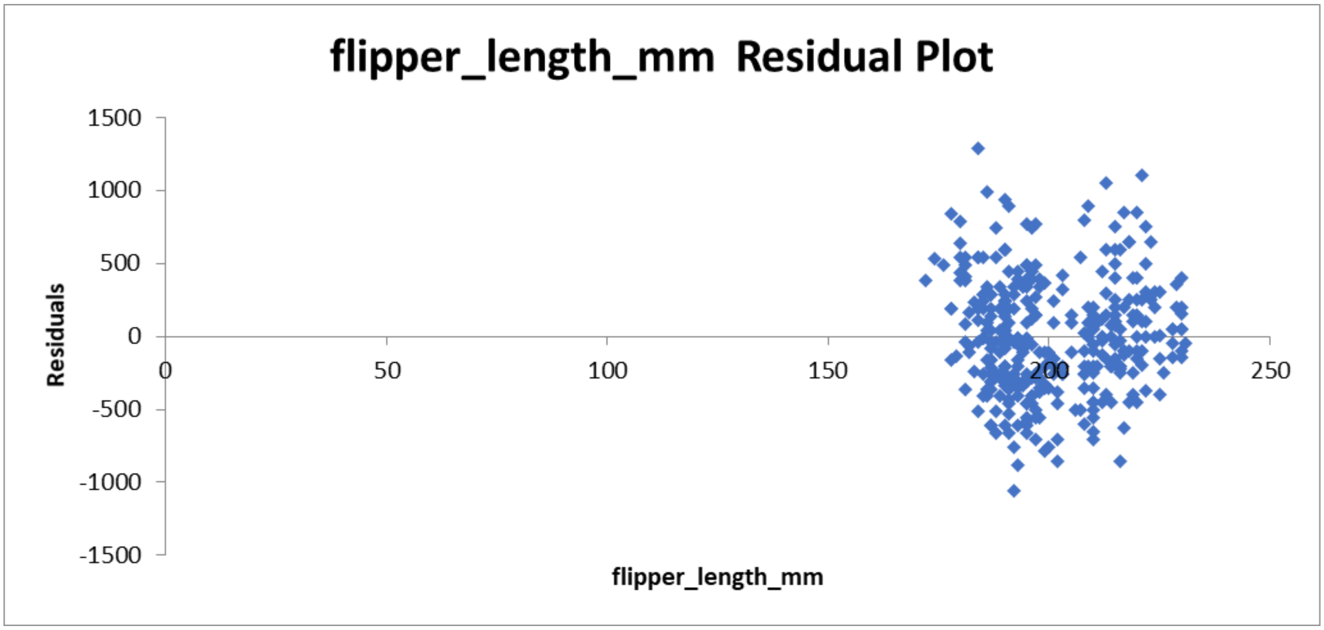



Drill: `penguins-linear.xlsx`

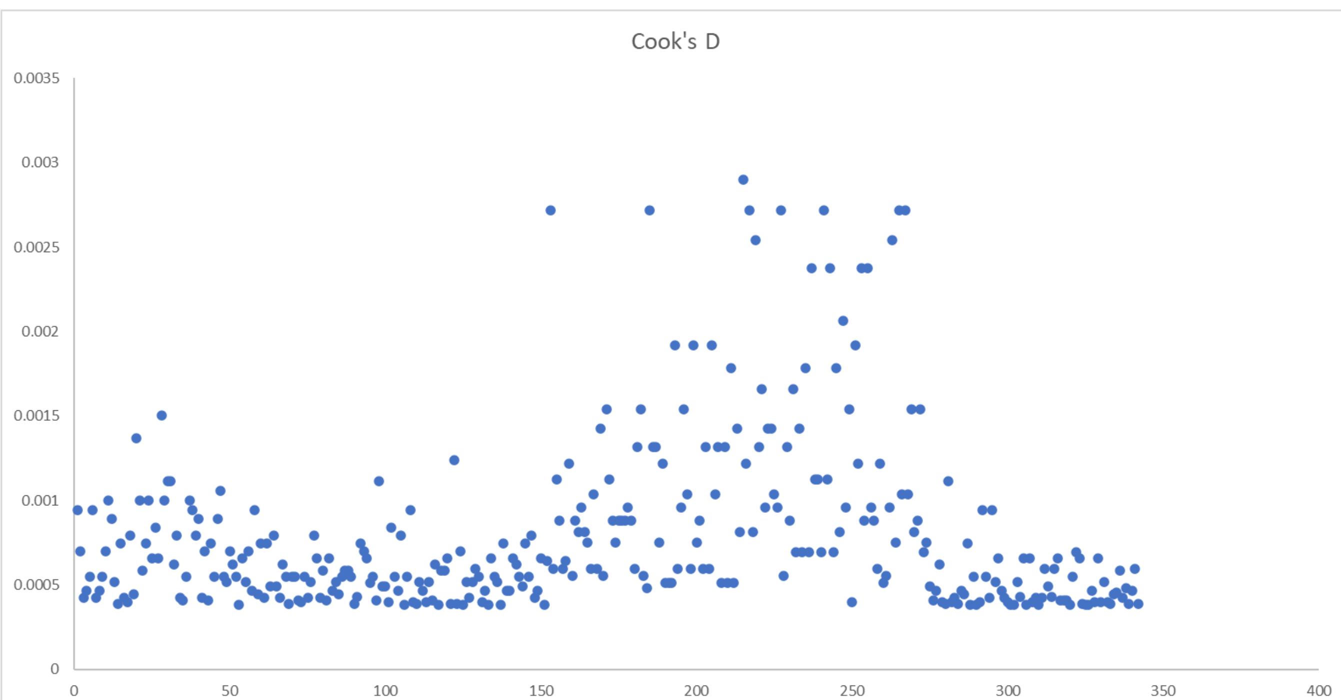
1. In this case, culmen length and culmen depth are not significant, so our only independent variable is flipper length. It's no longer multiple linear regression, but we can still check the same types of diagnostics.
   a. Linearity checks out from the scatter plot.




   b. There is no pattern in the residuals.




c. While none of the Cook's d values are over 1, it appears that some group of observations do have more influence over the curve than others. There are some ways to investigate that are outside the scope of this course.




Demo: `mpg-regression-diagnostics.xlsx`

1. Because this dataset uses multiple independent variables, we should use the *adjusted* r-square. That is available in the ToolPak results.
    a. An adjusted r-square means that 70% of the variance in Y is explained by our X's.

| | | |
|---|---|---|
| SUMMARY OUTPUT | | |
| | | |
| *Regression Statistics* | | |
| Multiple R | 0.840461346 | |
| R Square | 0.706375274 | |
| Adjusted R Square | 0.704865635 | <- 70% of the variance in Y is explained by X's |
| Standard Error | 4.240169468 | |
| Observations | 392 | |

2. We can also use the estimated coefficients from our output to make point predictions. Use the intercept and slopes to predict a value of Y given some X's:

| | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|
| 1 | | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | | |
| 3 | | *Regression Statistics* | | | | | | |
| 4 | | Multiple R | 0.840461346 | | | | | |
| 5 | | R Square | 0.706375274 | | | 200 | 3000 | |
| 6 | | Adjusted R Square | 0.704865635 | <- 70% of the variance in Y is explained by X's | | 18.79716613 | <- What is the expected MPG of a car weighing 3,000 pounds that has 200 horsepower? | |
| 7 | | Standard Error | 4.240169468 | | | =H17+(K5*H18)+(L5*H19) | | |
| 8 | | Observations | 392 | | | | | |
| 9 | | | | | | | | |
| 10 | | ANOVA | | | | | | |
| 11 | | | *df* | *SS* | *MS* | *F* | *Significance F* | |
| 12 | | Regression | 2 | 16825.14803 | 8412.574016 | 467.9101535 | 3.0596E-104 | |
| 13 | | Residual | 389 | 6993.845437 | 17.97903711 | | | |
| 14 | | Total | 391 | 23818.99347 | | | | |
| 15 | | | | | | | | |
| 16 | | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| 17 | | Intercept | 45.64021084 | 0.793195833 | 57.5396503 | 2.3171E-192 | 44.08072353 | 47.19969815 |
| 18 | | horsepower | -0.047302863 | 0.011085086 | -4.267252507 | 2.48848E-05 | -0.069097042 | -0.025508684 |
| 19 | | weight | -0.005794157 | 0.000502327 | -11.53463263 | 1.12436E-26 | -0.006781773 | -0.004806542 |
| 20 | | | | | | | | |
| 21 | | | | | | | | |

Drill: `penguin-linear-diagnostics.xlsx`

1. Follow the same steps as above. Notice that since this time we only have one independent variable, the r-square and adjusted r-square are the same.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | |
| 2 | | | | |
| 3 | *Regression Statistics* | | | |
| 4 | Multiple R | | 0.871201767 | |
| 5 | R Square | | 0.758992519 | <= 76% of variability in Y is caused by X |
| 6 | Adjusted R Square | | 0.758283674 | |
| 7 | Standard Error | | 394.2781775 | |
| 8 | Observations | | 342 | |
| 9 | | | | |

**Interaction terms**

Demo: `airquality-interaction.xlsx`

It's not a bad idea to run the regression *without* the interaction terms at first to establish a baseline.

1. We will set up our interaction term in column E by multiplying columns C and D.
2. For our baseline (the DV on the two IV's), about 50% of the variability in Y is explained.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | id | Ozone | Temp | Solar.R | temp * solar |
| 2 | 1 | 41 | 67 | 190 | 12730 |
| 3 | 2 | 36 | 72 | 118 | 8496 |
| 4 | 3 | 12 | 74 | 149 | 11026 |
| 5 | 4 | 18 | 62 | 313 | 19406 |
| 6 | 7 | 23 | 65 | 299 | 19435 |
| 7 | 8 | 19 | 59 | 99 | 5841 |
| 8 | 9 | 8 | 61 | 19 | 1159 |
| 9 | 12 | 16 | 69 | 256 | 17664 |
| 10 | 13 | 11 | 66 | 290 | 19140 |
| 11 | 14 | 14 | 68 | 274 | 18632 |
| 12 | 15 | 18 | 58 | 65 | 3770 |
| 13 | 16 | 14 | 64 | 334 | 21376 |
| 14 | 17 | 34 | 66 | 307 | 20262 |
| 15 | 18 | 6 | 57 | 78 | 4446 |
| 16 | 19 | 30 | 68 | 322 | 21896 |
| 17 | 20 | 11 | 62 | 44 | 2728 |
| 18 | 21 | 1 | 59 | 8 | 472 |
| 19 | 22 | 11 | 73 | 320 | 23360 |
| 20 | 23 | 4 | 61 | 25 | 1525 |
| 21 | 24 | 32 | 61 | 92 | 5612 |

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.71436457 |
| R Square | 0.510316738 |
| Adjusted R Square | 0.50124853 |
| Standard Error | 23.50026724 |
| Observations | 111 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 2 | 62157.5534 | 31078.7767 | 56.27536418 | 1.80063E-17 |
| Residual | 108 | 59644.35651 | 552.2625603 | | |
| Total | 110 | 121801.9099 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -145.7031551 | 18.44671758 | -7.898595208 | 2.52933E-12 | -182.2677495 | -109.1385607 | -182.2677495 | -109.1385607 |
| Temp | 2.278466835 | 0.24599582 | 9.262217671 | 2.21556E-15 | 1.790860443 | 2.766073227 | 1.790860443 | 2.766073227 |
| Solar.R | 0.057109594 | 0.025718852 | 2.220534321 | 0.028470633 | 0.006130367 | 0.10808882 | 0.006130367 | 0.10808882 |

3. If we add the interaction term, our adjusted R-square increases to .54. Is it worth adding an extra variable?

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 16 | 19 | 30 | 68 | 322 | 21896 |
| 17 | 20 | 11 | 62 | 44 | 2728 |
| 18 | 21 | 1 | 59 | 8 | 472 |
| 19 | 22 | 11 | 73 | 320 | 23360 |
| 20 | 23 | 4 | 61 | 25 | 1525 |
| 21 | 24 | 32 | 61 | 92 | 5612 |
| 22 | 28 | 23 | 67 | 13 | 871 |
| 23 | 29 | 45 | 81 | 252 | 20412 |
| 24 | 30 | 115 | 79 | 223 | 17617 |
| 25 | 31 | 37 | 76 | 279 | 21204 |
| 26 | 38 | 29 | 82 | 127 | 10414 |
| 27 | 40 | 71 | 90 | 291 | 26190 |
| 28 | 41 | 39 | 87 | 323 | 28101 |
| 29 | 44 | 23 | 82 | 148 | 12136 |
| 30 | 47 | 21 | 77 | 191 | 14707 |
| 31 | 48 | 37 | 72 | 284 | 20448 |
| 32 | 49 | 20 | 65 | 37 | 2405 |
| 33 | 50 | 12 | 73 | 120 | 8760 |
| 34 | 51 | 13 | 76 | 137 | 10412 |
| 35 | 62 | 135 | 84 | 269 | 22596 |
| 36 | 63 | 49 | 85 | 248 | 21080 |
| 37 | 64 | 32 | 81 | 236 | 19116 |
| 38 | 66 | 64 | 83 | 175 | 14525 |
| 39 | 67 | 40 | 83 | 314 | 26062 |
| 40 | 68 | 77 | 88 | 276 | 24288 |
| 41 | 69 | 97 | 92 | 267 | 24564 |
| 42 | 70 | 97 | 92 | 272 | 25024 |
| 43 | 71 | 85 | 89 | 175 | 15575 |
| 44 | 73 | 10 | 73 | 264 | 19272 |
| 45 | 74 | 27 | 81 | 175 | 14175 |
| 46 | 76 | 7 | 80 | 48 | 3840 |
| 47 | 77 | 48 | 81 | 260 | 21060 |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -145.7031551 | 18.44671758 | -7.898595208 | 2.52933E-12 | -182.2677495 | -109.1385607 | -182.2677495 | -109.1385607 |
| Temp | 2.278466835 | 0.24599582 | 9.262217671 | 2.21556E-15 | 1.790860443 | 2.766073227 | 1.790860443 | 2.766073227 |
| Solar.R | 0.057109594 | 0.025718852 | 2.220534321 | 0.028470633 | 0.006130367 | 0.10808882 | 0.006130367 | 0.10808882 |

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.742486766 |
| R Square | 0.551286598 |
| Adjusted R Square | 0.538705848 |
| Standard Error | 22.60058501 |
| Observations | 111 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 3 | 67147.76055 | 22382.58685 | 43.81985304 | 1.49244E-18 |
| Residual | 107 | 54654.14936 | 510.7864426 | | |
| Total | 110 | 121801.9099 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | -36.89906236 | 39.07007244 | -0.944432914 | 0.347076192 | -114.3509242 | 40.55279952 | -114.3509242 | 40.55279952 |
| Temp | 0.760655388 | 0.540163008 | 1.408196002 | 0.161971967 | -0.310154776 | 1.831465552 | -0.310154776 | 1.831465552 |
| Solar.R | -0.547794547 | 0.195103591 | -2.807711251 | 0.005931117 | -0.934564661 | -0.161024433 | -0.934564661 | -0.161024433 |
| temp * solar | 0.008274977 | 0.002647447 | 3.125644678 | 0.002284303 | 0.003026723 | 0.013523232 | 0.003026723 | 0.013523232 |

airquality

**Linear regression with categorical IV's**

**Demo:** `mpg-dummy.xlsx`

1. We want to model the influence of weight and origin on mpg. Because origin is categorical, we will create dummy variables to include as IV's.
   a. Origin takes three values: USA, Europe and Asia. We will encode Europe as one dummy variable, and Asia another. USA is implied when both dummies are set to zero. We will not include this column in our regression since it's all zeros.

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | =IF(F3="Europe",1,0) | =IF(F3="Asia",1,0) | =IF(F3="USA",0,0) | | |
| 2 | mpg | weight | origin_europe | origin_asia | origin_usa | origin | car name |
| 3 | 18 | 3504 | 0 | 0 | 0 | USA | chevrolet chevelle malibu |
| 4 | 15 | 3693 | 0 | 0 | 0 | USA | buick skylark 320 |
| 5 | 18 | 3436 | 0 | 0 | 0 | USA | plymouth satellite |
| 6 | 16 | 3433 | 0 | 0 | 0 | USA | amc rebel sst |
| 7 | 17 | 3449 | 0 | 0 | 0 | USA | ford torino |
| 8 | 15 | 4341 | 0 | 0 | 0 | USA | ford galaxie 500 |
| 9 | 14 | 4354 | 0 | 0 | 0 | USA | chevrolet impala |
| 10 | 14 | 4312 | 0 | 0 | 0 | USA | plymouth fury iii |
| 11 | 14 | 4425 | 0 | 0 | 0 | USA | pontiac catalina |
| 12 | 15 | 3850 | 0 | 0 | 0 | USA | amc ambassador dpl |
| 13 | 15 | 3563 | 0 | 0 | 0 | USA | dodge challenger se |
| 14 | 14 | 3609 | 0 | 0 | 0 | USA | plymouth 'cuda 340 |

2. Run the regression from the ToolPak, exporting the results to cell K1 of the same worksheet. You will get the following results.
    a. These value give the coefficients to include with our dummy-coded variables.
        i. Because the dummy-code for USA was left to zero, this becomes our reference category. We see here from the p-values that European cars do not have a significantly higher mileage from American cars, but Japanese cars do.
            1. However, we cannot keep some dummy variables and drop others. That would lead to inaccurate comparisons across groups. Since this p-value is so close to .05, I will decide to keep it in the model.

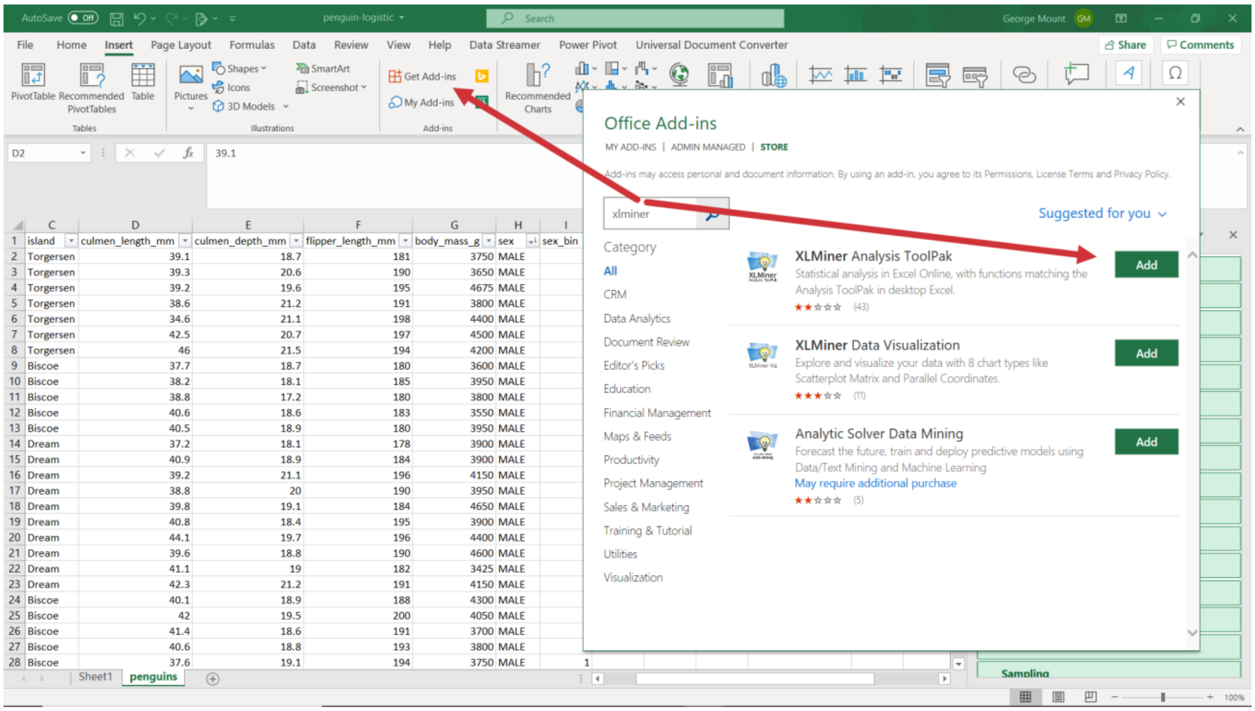
| | K | L | M | N | O | P | Q | R | S |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | | | |
| 2 | | | | | | Weight | origin_europe | origin_asia | mpg_pred |
| 3 | *Regression Statistics* | | | | American | 3000 | 0 | 0 | |
| 4 | Multiple R | 0.837558007 | | | European | 3000 | 1 | 0 | |
| 5 | R Square | 0.701503415 | | | Asian | 3000 | 0 | 1 | |
| 6 | Adjusted R Square | 0.699230599 | | | | | | | |
| 7 | Standard Error | 4.286477069 | | | | | | | |
| 8 | Observations | 398 | | | | | | | |
| 9 | | | | | | | | | |
| 10 | ANOVA | | | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| 12 | Regression | 3 | 17013.26452 | 5671.088175 | 308.6493667 | 4.8646E-103 | | | |
| 13 | Residual | 394 | 7239.310953 | 18.37388567 | | | | | |
| 14 | Total | 397 | 24252.57548 | | | | | | |
| 15 | | | | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| 17 | Intercept | 43.69585641 | 1.104363368 | 39.56655724 | 2.5821E-139 | 41.5246745 | 45.86703833 | 41.5246745 | 45.86703833 |
| 18 | weight | -0.007023439 | 0.000318398 | -22.05865536 | 8.40858E-71 | -0.007649411 | -0.006397467 | -0.007649411 | -0.006397467 |
| 19 | origin_europe | 1.215471794 | 0.652373629 | 1.863152862 | 0.063184619 | -0.067096849 | 2.498040436 | -0.067096849 | 2.498040436 |
| 20 | origin_asia | 2.355434709 | 0.662030569 | 3.557894182 | 0.000419392 | 1.053880491 | 3.656988926 | 1.053880491 | 3.656988926 |
| 21 | | | | | | | | | |

3. We can now use these coefficients to make point predictions in the range O2:S5. Notice that we could technically include each of the dummy-coded variables in our equations; but the non-necessary coefficients are multiplied by zero.

| | K | L | M | N | O | P | Q | R | S | T | U | V | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | | | | | | | |
| 2 | | | | | | Weight | origin_europe | origin_asia | mpg_pred | | | | |
| 3 | *Regression Statistics* | | | | American | 3000 | 0 | 0 | 22.62553953 | =L17+((P3*$L$18)+(Q3*L19)+(R3*L20)) | | | |
| 4 | Multiple R | 0.837558007 | | | European | 3000 | 1 | 0 | 23.84101132 | =L17+(P4*$L$18)+(Q4*L19)+(R4*L20) | | | |
| 5 | R Square | 0.701503415 | | | Asian | 3000 | 0 | 1 | 24.98097424 | =L17+(P5*$L$18)+(L19*Q5)+(R5*$L$20) | | | |
| 6 | Adjusted R Square | 0.699230599 | | | | | | | | | | | |
| 7 | Standard Error | 4.286477069 | | | | | | | | | | | |
| 8 | Observations | 398 | | | | | | | | | | | |
| 9 | | | | | | | | | | | | | |
| 10 | ANOVA | | | | | | | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | | | | | |
| 12 | Regression | 3 | 17013.26452 | 5671.088175 | 308.6493667 | 4.8646E-103 | | | | | | | |
| 13 | Residual | 394 | 7239.310953 | 18.37388567 | | | | | | | | | |
| 14 | Total | 397 | 24252.57548 | | | | | | | | | | |
| 15 | | | | | | | | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* | | | | |
| 17 | Intercept | 43.69585641 | 1.104363368 | 39.56655724 | 2.5821E-139 | 41.5246745 | 45.86703833 | 41.5246745 | 45.86703833 | | | | |
| 18 | weight | -0.007023439 | 0.000318398 | -22.05865536 | 8.40858E-71 | -0.007649411 | -0.006397467 | -0.007649411 | -0.006397467 | | | | |
| 19 | origin_europe | 1.215471794 | 0.652373629 | 1.863152862 | 0.063184619 | -0.067096849 | 2.498040436 | -0.067096849 | 2.498040436 | | | | |
| 20 | origin_asia | 2.355434709 | 0.662030569 | 3.557894182 | 0.000419392 | 1.053880491 | 3.656988926 | 1.053880491 | 3.656988926 | | | | |
| 21 | | | | | | | | | | | | | |
| 22 | | | | | | | | | | | | | |

**Logistic regression**

0. The Analysis ToolPak does not include logistic regression, so you need to install XLMiner. Fortunately this is free and requires no external downloads: go to **Insert > Get Add-Ins > XLMiner Analysis ToolPak.**




## Demo: `occupancy.xlsx`

1. Select your Y range (`Occupancy`) and X range (`Temperature`, `Humidity`, and `Light`).
   a. The range selector tool in XLMiner is terribly user-unfriendly. It may be easier to grab a smaller range and fill it out by typing.
2. We will get some familiar output expressed in intercepts and p-values. All of our X's are significant.

| | H | I | J | K | L | M | N | O | P | Q |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | | | | |
| 2 | | | | | | | | | | |
| 3 | Regression Statistics | | | | | | | | | |
| 4 | Chi Square | 7305.64272 | | | | | | | | |
| 5 | Residual Dev. | 1114.657625 | | | | | | | | |
| 6 | # of iterations | 10 | | | | | | | | |
| 7 | Observations | 8143 | | | | | | | | |
| 8 | | | | | | | | | | |
| 9 | | Coefficients | Standard Error | P-value | Odd Ratio | Lower 95% | Upper 95% | Lower 95% | Upper 95% | |
| 10 | Intercept | 0.47203998 | 2.337356777 | 0.839952208 | 1.603261 | 0.016423 | 156.5186 | 0.016423 | 156.5186 | |
| 11 | Temperature | -0.586523047 | 0.110473391 | 1.1012E-07 | 0.556258 | 0.447962 | 0.690735 | 0.447962 | 0.690735 | |
| 12 | Humidity | 0.145701588 | 0.017123159 | 1.75391E-17 | 1.156851 | 1.11867 | 1.196335 | 1.11867 | 1.196335 | |
| 13 | Light | 0.024478643 | 0.0009229 | 5.179E-155 | 1.024781 | 1.022929 | 1.026636 | 1.022929 | 1.026636 | |
| 14 | | | | | | | | | | |

Drill: `penguin-logistic.xlsx`

1. XLMiner requires that all variables be *numeric.* This means that the `MALE`/`FEMALE` labels for `sex` should be converted into 0's and 1's.

I2 =IF(H2="MALE",0,1)

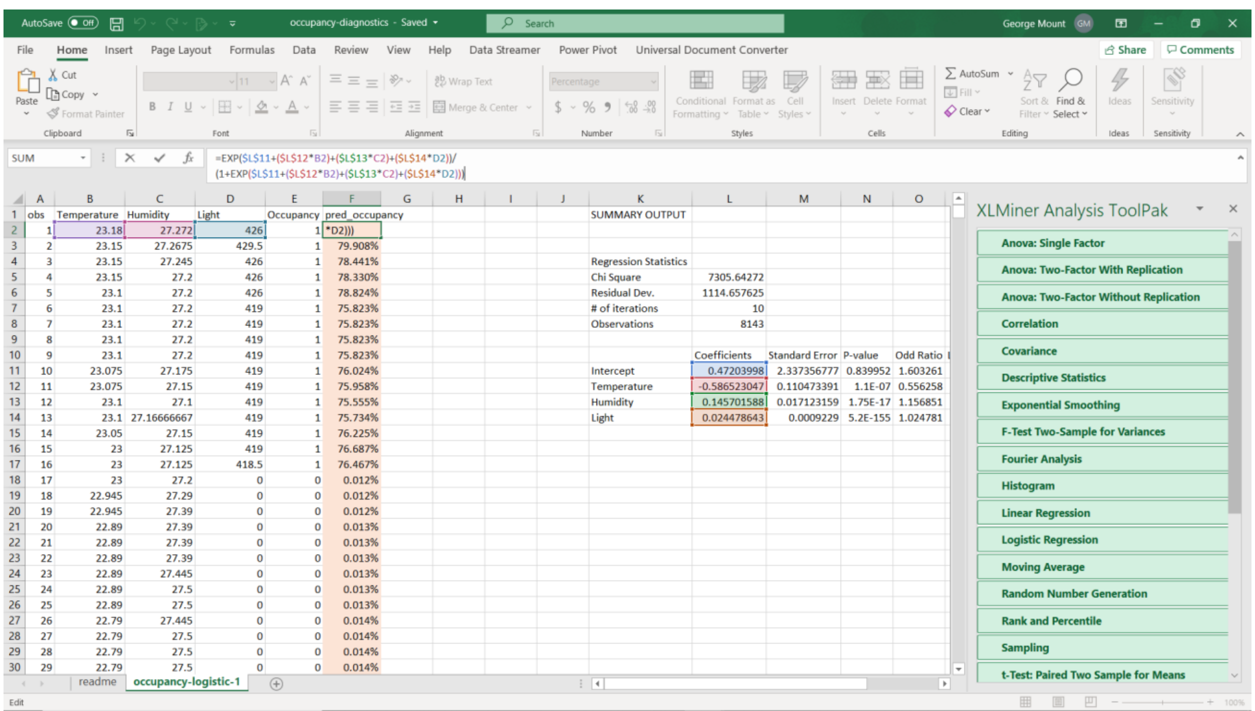
| | D | E | F | G | H | I |
|---|---|---|---|---|---|---|
| 1 | culmen_length_mm | culmen_depth_mm | flipper_length_mm | body_mass_g | sex | sex_bin |
| 2 | 39.1 | 18.7 | 181 | 3750 | MALE | 0 |
| 3 | 39.3 | 20.6 | 190 | 3650 | MALE | 0 |
| 4 | 39.2 | 19.6 | 195 | 4675 | MALE | 0 |
| 5 | 38.6 | 21.2 | 191 | 3800 | MALE | 0 |
| 6 | 34.6 | 21.1 | 198 | 4400 | MALE | 0 |
| 7 | 42.5 | 20.7 | 197 | 4500 | MALE | 0 |
| 8 | 46 | 21.5 | 194 | 4200 | MALE | 0 |

2. `flipper_length_mm` is not significant, so drop it and re-run.

| | K | L | M | N | O | P | Q | R | S | T | U |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | | | | | |
| 2 | | | | | | | | | | | |
| 3 | Regression Statistics | | | | | | | | | | |
| 4 | Chi Square | 302.6055688 | | | | | | | | | |
| 5 | Residual Dev. | 159.0034261 | | | | | | | | | |
| 6 | # of iterations | 8 | | | | | | | | | |
| 7 | Observations | 333 | | | | | | | | | |
| 8 | | | | | | | | | | | |
| 9 | | Coefficients | Standard Error | P-value | Odd Ratio | Lower 95% | Upper 95% | Lower 95% | Upper 95% | | |
| 10 | Intercept | 56.11740399 | 8.369320872 | 2.01227E-11 | 2.35223E+24 | 1.77E+17 | 3.13E+31 | 1.77E+17 | 3.13E+31 | | |
| 11 | culmen_length_mm | -0.107629548 | 0.047967294 | 0.024844556 | 0.897960186 | 0.817386 | 0.986477 | 0.817386 | 0.986477 | | |
| 12 | culmen_depth_mm | -2.031515596 | 0.249465002 | 3.84064E-16 | 0.13113662 | 0.080423 | 0.21383 | 0.080423 | 0.21383 | | |
| 13 | flipper_length_mm | 0.032474395 | 0.034755396 | 0.350113129 | 1.033007443 | 0.964983 | 1.105827 | 0.964983 | 1.105827 | | |
| 14 | body_mass_g | -0.005512026 | 0.000823488 | 2.17886E-11 | 0.994503137 | 0.992899 | 0.99611 | 0.992899 | 0.99611 | | |
| 15 | | | | | | | | | | | |
| 16 | | | | | | | | | | | |
| 17 | | | | | | | | | | | |

Demo: `occupancy-diagnostics.xlsx`

1. The equation to find the probability is a doozy, but this comes from the logit equation. Read through it to interpret!




2. We can assume that any probability greater than 50% is a "yes," otherwise "no." Find this predicted outcome in column G.

G2 =IF(F2>0.5,1,0)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | obs | Temperature | Humidity | Light | Occupancy | pred_occupancy | pred_occupied? | pred_correct? | |
| 2 | 1 | 23.18 | 27.272 | 426 | 1 | 78.209% | 1 | TRUE | |
| 3 | 2 | 23.15 | 27.2675 | 429.5 | 1 | 79.908% | 1 | TRUE | |
| 4 | 3 | 23.15 | 27.245 | 426 | 1 | 78.441% | 1 | TRUE | |
| 5 | 4 | 23.15 | 27.2 | 426 | 1 | 78.330% | 1 | TRUE | |
| 6 | 5 | 23.1 | 27.2 | 426 | 1 | 78.824% | 1 | TRUE | |
| 7 | 6 | 23.1 | 27.2 | 419 | 1 | 75.823% | 1 | TRUE | |

3. Now we can simply find a TRUE/FALSE result as to whether the predicted outcome is the same as the actual one.

H2 =E2=G2

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | obs | Temperature | Humidity | Light | Occupancy | pred_occupancy | pred_occupied? | pred_correct? | |
| 2 | 1 | 23.18 | 27.272 | 426 | 1 | 78.209% | 1 | TRUE | |
| 3 | 2 | 23.15 | 27.2675 | 429.5 | 1 | 79.908% | 1 | TRUE | |
| 4 | 3 | 23.15 | 27.245 | 426 | 1 | 78.441% | 1 | TRUE | |
| 5 | 4 | 23.15 | 27.2 | 426 | 1 | 78.330% | 1 | TRUE | |
| 6 | 5 | 23.1 | 27.2 | 426 | 1 | 78.824% | 1 | TRUE | |
| 7 | 6 | 23.1 | 27.2 | 419 | 1 | 75.823% | 1 | TRUE | |
| 8 | 7 | 23.1 | 27.2 | 419 | 1 | 75.823% | 1 | TRUE | |

4. We can now calculate a basic predictive accuracy measure for this model. 98.7% isn't too bad for a first pass!

| | K | L | M | N | O | P |
|---|---|---|---|---|---|---|
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | Number observations | 8143 | =COUNT(A2:A8144) | | | |
| 17 | Number observations predicted correctly? | 8039 | =COUNTIF(H2:H8144,"TRUE") | | | |
| 18 | Pred. accuracy | 98.7% | =L17/L16 | | | |
| 19 | | | | | | |

5. Now that we have set up our equations in the table, we can easily plug in any point-estimates to predict a "yes" or "no" outcomes

a. For temperature 25, humidity 30 and light 400, it's a close call! Maybe making binary predictions isn't so straightforward after all.

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | obs | Temperature | Humidity | Light | Occupancy | pred_occupancy | pred_occupied? | pred_correct? | |
| 2 | | 25 | 30 | 400 | | 49.287% | 0 | | |
| 3 | 1 | 23.18 | 27.272 | 426 | 1 | 78.209% | 1 | TRUE | |
| 4 | 2 | 23.15 | 27.2675 | 429.5 | 1 | 79.908% | 1 | TRUE | |
| 5 | 3 | 23.15 | 27.245 | 426 | 1 | 78.441% | 1 | TRUE | |

## Drill: `penguin-logistic-diagnostics.xlsx`

1. Follow the same steps as above. See if you can write the logit curve equation on your own!