## Regression validation

The topic of model validation can be divided into two parts:

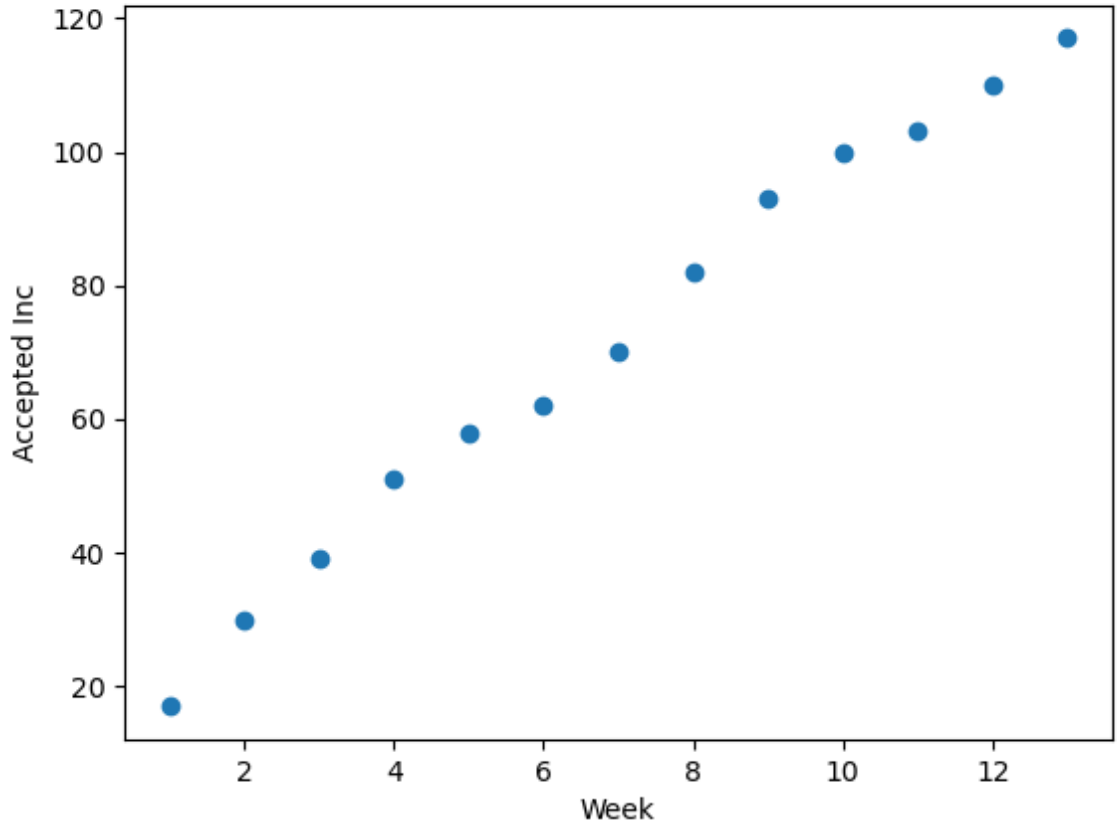
1. How well the model fits the training data (and how much it is wrong)

2. What capacity it has to predict observations that are not present in the training data. This topic was covered in the validation chapter for the 3D model we discussed, but I understand that the concepts can be explained in 2D to understand them and then extend the reasoning to more dimensions.

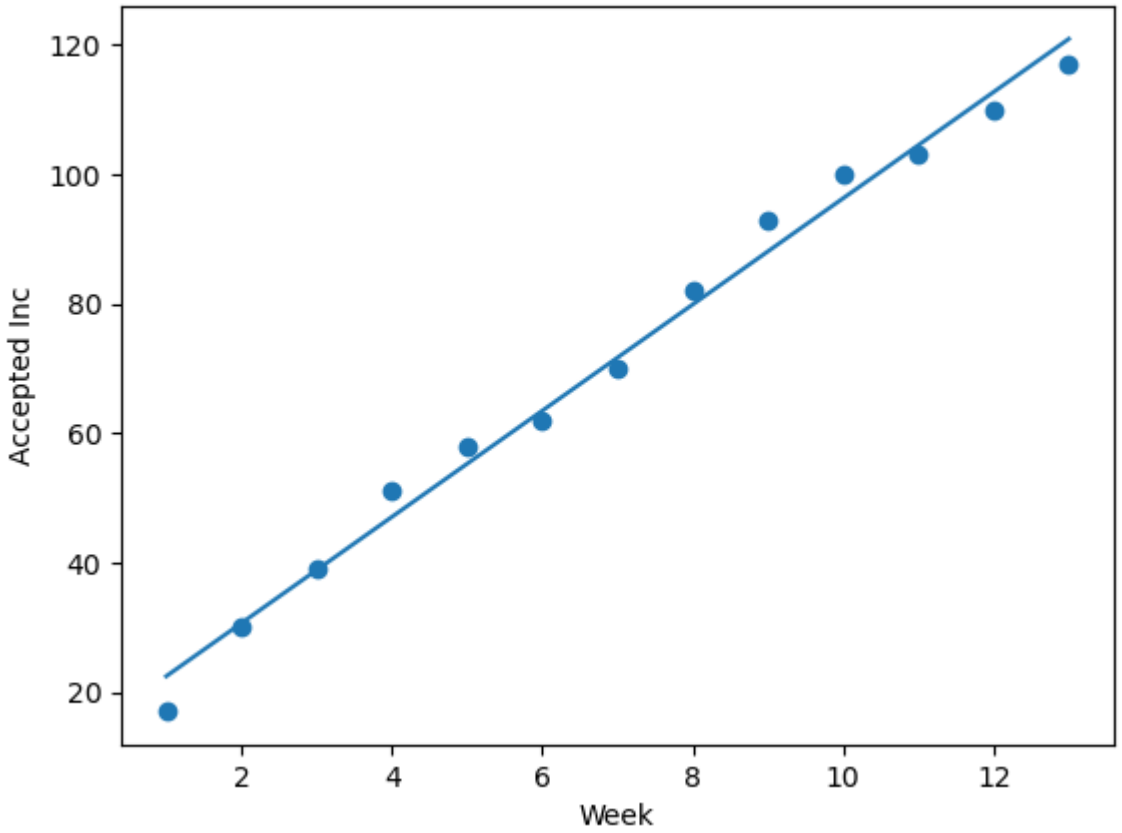
To explain the simplified model, let's take a series of example data for campus TPA, term 2022FAQ, program ADN.2YR.AS.

| Accepted Inc | week |
|---|---|
| 17 | 1 |
| 30 | 2 |
| 39 | 3 |
| 51 | 4 |
| 58 | 5 |
| 62 | 6 |
| 70 | 7 |
| 82 | 8 |
| 93 | 9 |
| 100 | 10 |
| 103 | 11 |
| 110 | 12 |
| 117 | 13 |

In the following graph you can see the evolution of accepted incomplete over time:

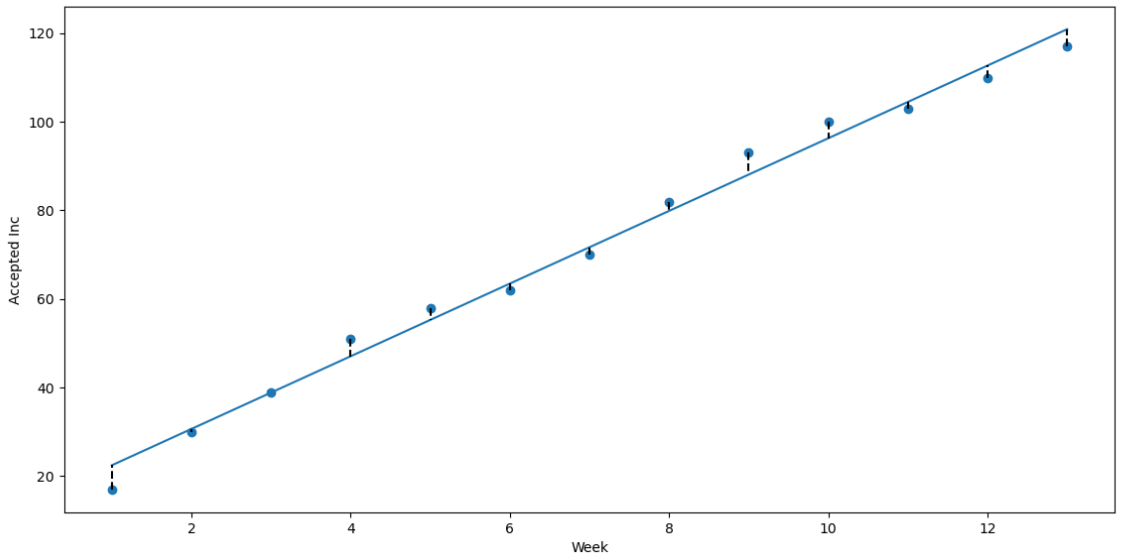



A linear regression is a line that fits quite well to the points as seen in this graph:



# Regarding topic 1: How well the model fits the training data (and how much it is wrong)

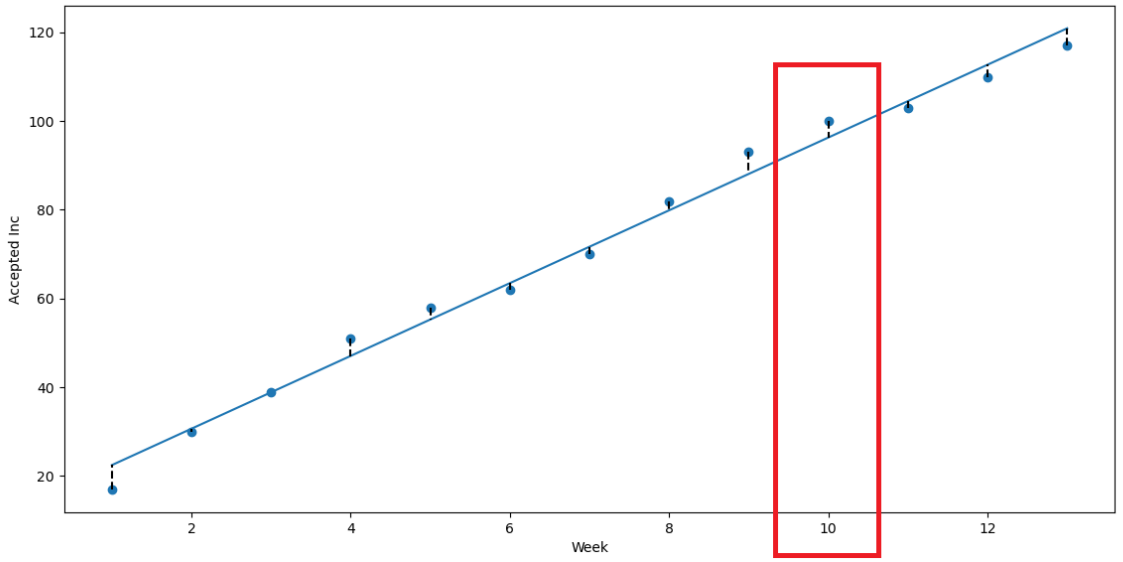
You can see graphically that the line passes close to the points but not exactly (black dotted lines).




This difference between what is predicted and the training data is used to calculate the estimation error.

The errors are averaged and a measure of how close or far away the model is from the training points is obtained (the error indicator is root mean squared error -RMSE-, the lower, the better).

Let's take as an example that we want to predict any week within the line, in this case, week 10.



In numbers it can be seen in this example:

| week | predicted | training value | error |
|---|---|---|---|
| 1 | 23 | 17 | 5,5 |
| 2 | 32 | 30 | 0,7 |
| 3 | 39 | 39 | -0,1 |
| 4 | 48 | 51 | -3,9 |
| 5 | 56 | 58 | -2,7 |
| 6 | 64 | 62 | 1,5 |
| 7 | 73 | 70 | 1,7 |
| 8 | 81 | 82 | -2,1 |
| 9 | 89 | 93 | -4,9 |
| 10 | 97 | 100 | -3,7 |
| 11 | 106 | 103 | 1,5 |
| 12 | 114 | 110 | 2,7 |
| 13 | 122 | 117 | 3,9 |

As can be seen in the table, for week 10, approximately 97 incomplete were estimated when in reality the training data is 100. This gives an error of -3.7 (taking into account the decimals).

The error can be due to overshooting or falling short, so the error can be negative or positive. To average them, they are squared, averaged and with the average a square root is applied to return to a value on the scale of what is being measured.

In this example:

| week | predicted | training value | error | error^2 |
|---|---|---|---|---|
| 1 | 23 | 17 | 5,5 | 29,9 |
| 2 | 32 | 30 | 0,7 | 0,5 |
| 3 | 39 | 39 | -0,1 | 0,0 |
| 4 | 48 | 51 | -3,9 | 15,2 |
| 5 | 56 | 58 | -2,7 | 7,3 |
| 6 | 64 | 62 | 1,5 | 2,3 |
| 7 | 73 | 70 | 1,7 | 2,9 |
| 8 | 81 | 82 | -2,1 | 4,4 |
| 9 | 89 | 93 | -4,9 | 24,0 |
| 10 | 97 | 100 | -3,7 | 13,7 |

| | | | | |
|---|---|---|---|---|
| 11 | 106 | 103 | 1,5 | 2,3 |
| 12 | 114 | 110 | 2,7 | 7,3 |
| 13 | 122 | 117 | 3,9 | 15,2 |
| | | | | |
| | | | Average of error^2 | 9,6 |
| | | | sqrt of avg (RMSE) | 3,1 |
| | | | | |

If the RMSE of a model is 3.1, this means that on average, the predictions of the model deviate from the actual values by 3.1 units. This does not necessarily mean that the model “makes a mistake” of 3 units in each prediction, as the RMSE is a measure of the dispersion of the errors.

The training RMSE measures how well the model fits the historical training data (the data used to adjust the regression line). However, this does not mean that the model will accurately predict new data that it has not been trained on. How to validate and measure the model is part of topic 2.

## Regarding topic 2: What capacity it has to predict observations that are not present in the training data.

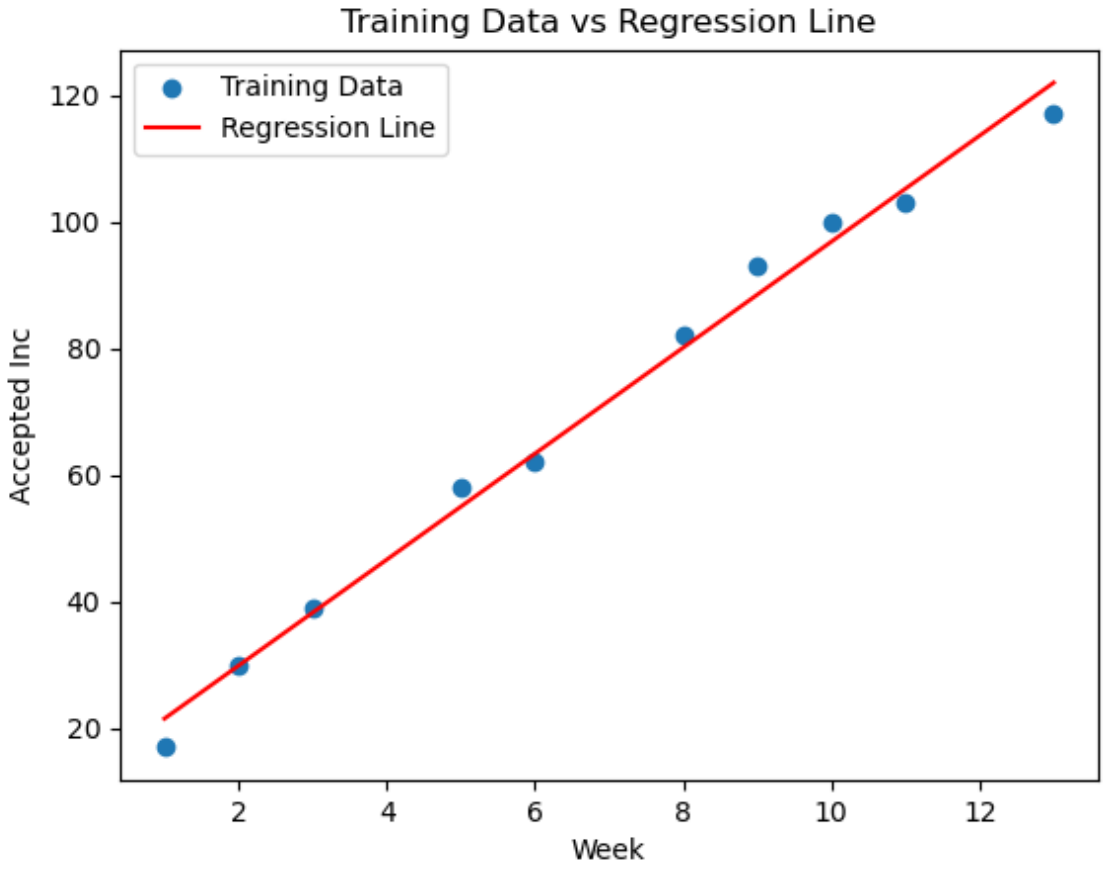
To measure the ability to generalize predictions to values not yet seen, the technique used is to divide the historical data into two groups.

A larger first set is used to perform the regression, in this case we will arbitrarily choose most of the points except 3 for didactic purposes. The training set will contain all weeks except 4, 7 and 12, so the training data looks like this:

| week | training value |
|---|---|
| 1 | 17 |
| 2 | 30 |
| 3 | 39 |
| | |
| 5 | 58 |
| 6 | 62 |
| | |
| 8 | 82 |
| 9 | 93 |

| | |
|---|---|
| 10 | 100 |
| 11 | 103 |
| | |
| 13 | 117 |

The regression of the training data looks like this graphically (without the points for weeks 4, 7, and 12).




In this particular case, using the training data, the regression adopts its final form that no longer changes with new points:

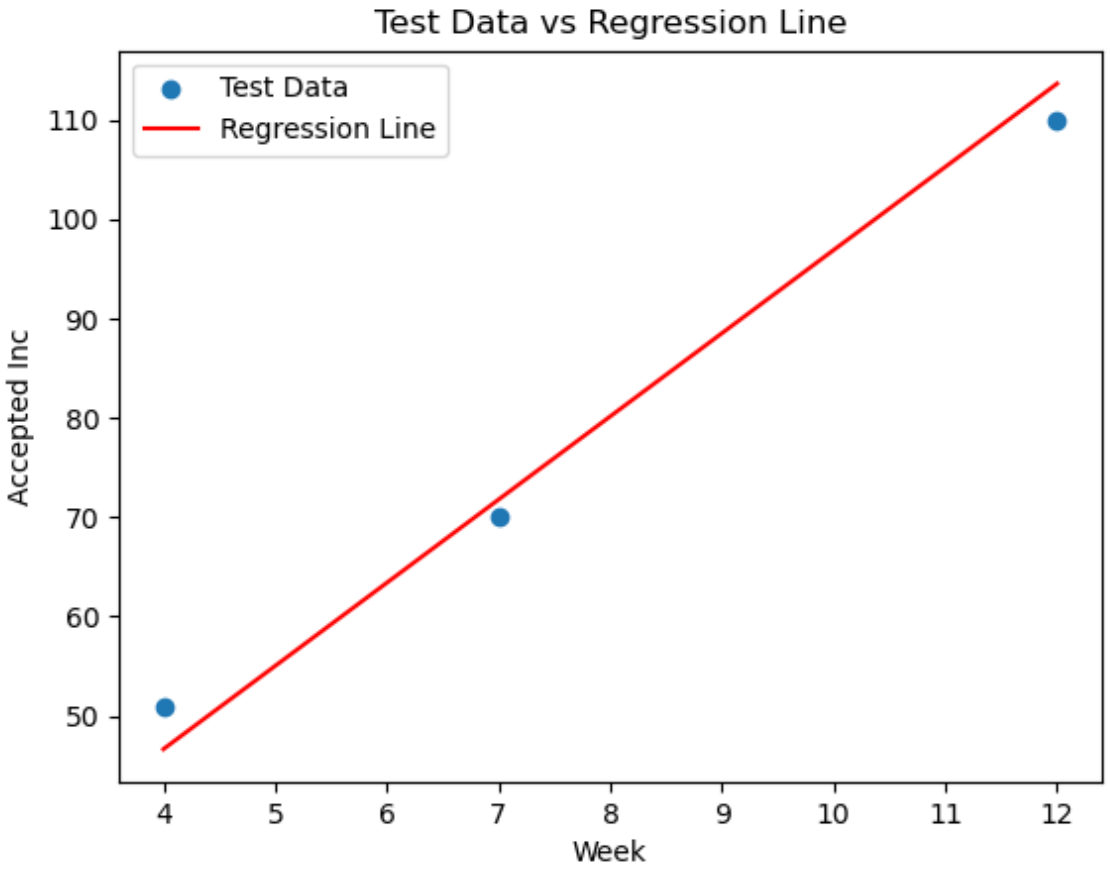
Linear Regression Formula: y = 8.38x + 13.15

where y is the accepted incomplete and x is weeks. So from now on, x can be replaced with other values to see what y predicts.

This is where the second set called the test set comes into play, which was not used for training, in this case weeks 4, 7 and 12:

| week | test value | t |
|---|---|---|
| 4 | 51 | |
| 7 | 70 | |
| 12 | 110 | |

Here you can see the test data not used for training, with respect to the regression:




The estimation is made for the test data, the error is calculated as in case 1, but on the test data, it is averaged and the test RMSE is calculated, which measures the ability to generalize to new data:

| week | real data value | estimation | error | error^2 |
|---|---|---|---|---|
| 4 | 51 | 46,6 | -4,4 | 18,9 |
| 7 | 70 | 71,8 | 1,8 | 3,2 |
| 12 | 110 | 113,7 | 3,7 | 13,3 |
| | | | | |
| | | | avg error^2 | 11,8 |
| | | | RMSE test | 3,44 |

The test RMSE is one of the main variables that is aimed to be minimized because it reduces the error of predicting new values.

In this case, the test RMSE is 3.44, which means that the prediction has a deviation of that value. These values must be considered with respect to business values to decide if these values are adequate, or how much RMSE is acceptable at a decisional level.

In the case of the work we are carrying out, the test RMSE is measured for each campus/program combination and I put it in the documentation (in the validation section) and also cited some cases with fancy or “unusual” RMSE numbers that, either because they are low or high, may represent some special or different behavior that may be worth analyzing to fine-tune the model to something more sophisticated.