In this knowledge check, you will get the chance to practice reading and interpreting regression outputs using Python.

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	16 minutes	5 out of 5

Answers will be shown after your last attempt

Score for this attempt: **5** out of 5
Submitted May 3 at 12:29am
This attempt took 16 minutes.

Question 1

1 / 1 pts

Upon examining the test (out-of-sample) evaluation metrics, you notice that the model seems to perform extremely poorly in comparison to training (in-sample) data, yet the large number of independent variables are all significant. What could the problem be?

- None of these answers
- The model is overfitting — there are too many variables.
- The sample size is too large, leading to increased variability.
- An outlier is disproportionately influencing the model.

That is correct! Overfitting can occur when there are too many variables in the model, leading to a model that fits well to the training data but not the test data. Some variables may be modeling noise.

Question 2

1 / 1 pts

The number of days required to manufacture a certain engine is distributed normally with a mean of 40 days and a standard deviation of 10 days. In Google Colab, you will need to `import scipy.stats as norm` before applying the formula, as demonstrated below:

```
from scipy.stats import norm
```

Find the probability that an engine will be completed in fewer than 50 days.

- 0.63
- 0.11
- 0.84
- 0.75

That is correct! This represents the area under the normal distribution curve, to the left of 50 days.

```
from scipy.stats import norm

norm.cdf(50, loc=40, scale=10)
```

Question 3

1 / 1 pts

Using the same scenario as Question 2, find the probability that an engine will be completed in more than 50 days.

- 0.22
- 0.38
- 0.25
- 0.16

That is correct! This represents the area under the normal distribution curve, to the right of 50 days.

```
from scipy.stats import norm

1 - norm.cdf(50, loc= 40, scale=10)
```

Question 4

1 / 1 pts

A model was run using training and test (out-of-sample) datasets. The following output displays the results.

What observations can you make about the data based on the output? **Select all that apply.**

```
Mean absolute error (train): 49.38
Mean absolute error (test) : 53.47

Mean squared error (train): 4,015.26
Mean squared error (test) : 4,083.26

Root mean squared error (train): 63.37
```

```
Root mean squared error (test) : 63.90

regressor.score (X_train, Y_train): 0.706878
regressor.score (X_test, Y_test)  : 0.391366
```

- [ ] The mean absolute error for the test (out-of-sample) dataset (53.47) is better than that of the training dataset (49.38).

- [x] The R-squared (R2) for the test (out-of-sample) dataset (0.391) is significantly worse than that of the training dataset (0.707).

- [x] The mean absolute error for the test (out-of-sample) dataset (53.47) is worse than that of the training dataset (49.38).

- [ ] The R-squared (R2) for the test (out-of-sample) dataset (0.391) is significantly better than that of the training dataset (0.707).

That is correct! The R-squared (R2) for the test (out-of-sample) dataset (0.391) is significantly worse than that of the training dataset (0.707) an the mean absolute error for the test (out-of-sample) dataset (53.47) is worse than that of the training dataset (49.38).

Question 5

1 / 1 pts

Using the same training and test (out-of-sample) datasets from Question 4 and based on your understanding of the data, is the model overfitting?

```
Mean absolute error (train): 49.38
Mean absolute error (test) : 53.47
```

```
Mean squared error (train): 4,015.26
Mean squared error (test) : 4,083.26

Root mean squared error (train): 63.37
Root mean squared error (test) : 63.90

regressor.score (X_train, Y_train): 0.706878
regressor.score (X_test, Y_test)  : 0.391366
```

- The model is not overfitting, as the R-squared (R2) for the test (out-of-sample) dataset (0.391) is significantly worse than that of the training dataset (0.707).

- (selected) The model is overfitting, as the R-squared (R2) for the test (out-of-sample) dataset (0.391) is significantly better than that of the training dataset (0.707).

- There is not enough data to make this determination.

That is correct! A worse R-squared in the test dataset indicates the model will perform poorly with new data. The model may be overfitting and modeling random noise in the training dataset.

Quiz Score: **5** out of 5