Regression exercises

Exercise 1

Predicting numeric values with Regression

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Predicting numeric values with Regression

Exercise 3.1

The t-statistics computed on Table 3.4 are computed individually for each coefficient since they are independent variables. Accordingly, there are 4 null hypotheses that we are testing:

1. H_0 for "TV": **in the presence of** Radio and Newspaper ads (and in addition to the intercept), there is no relationship between TV and Sales;
2. H_0 for "Radio": **in the presence of** TV and Newspaper ads (and in addition to the intercept), there is no relationship between Radio and Sales;
3. H_0 for "Newspaper": **in the presence of** TV and Radio ads (and in addition to the intercept), there is no relationship between Newspaper and Sales;
4. H_0 for the intercept: **in the absence of** TV, Radio and Newspaper ads, Sales are zero;

versus the 4 corresponding alternative hypotheses:

H_a: There is some relationship between TV/Radio/Newspaper and Sales, or Sales are non-zero in the absence of the other variables.

https://botlnec.github.io/islp/sols/chapter3/exercise1/

Predicting numeric values with Regression

3. Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 = \text{Interaction between GPA and IQ}$, and $X_5 = \text{Interaction between GPA and Gender}$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Predicting numeric values with Regression

Exercise 3.3

(a)

$$Y = \beta_0 + \beta_1 \times GPA + \beta_2 \times IQ + \beta_3 \times Gender + \beta_4 \times GPA \times IQ + \beta_5 \times GPA \times Gender$$

For a fixed value of GPA and IQ, the difference between female and male is given by:

$$Y_{female} - Y_{male} = \beta_3 + \beta_5 \times GPA = 35 - 10GPA,$$

which depends on GPA. It is clear that in the normal range of the GPA (0 to 4.0), the difference in expected salary between female and male ranges linearly from 35 to -5. In particular, if GPA > 3.5, males earn on average more than females. Therefore, **the correct answer is (iii)**.

https://botlnec.github.io/islp/sols/chapter3/exercise3/

Predicting numeric values with Regression

8. This question involves the use of simple linear regression on the `Auto` data set.

   (a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

   i. Is there a relationship between the predictor and the response?
   ii. How strong is the relationship between the predictor and the response?
   iii. Is the relationship between the predictor and the response positive or negative?
   iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95 % confidence and prediction intervals?

   (b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

   (c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Predicting numeric values with Regression

Exercise 3.8

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf  # Statsmodels is a Python module for statistics
import statsmodels.api as sm

%matplotlib inline
```

```
df = pd.read_csv('../data/auto.csv') # import dataset
```

```
df.head() # just to have a look
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

https://botlnec.github.io/islp/sols/chapter3/exercise8/
http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf | Chapter 3

Exercise 2

Predicting numeric values with Regression

1. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Predicting numeric values with Regression

Exercise 3.1

The t-statistics computed on Table 3.4 are computed individually for each coefficient since they are independent variables. Accordingly, there are 4 null hypotheses that we are testing:

1. H_0 for "TV": **in the presence of** Radio and Newspaper ads (and in addition to the intercept), there is no relationship between TV and Sales;
2. H_0 for "Radio": **in the presence of** TV and Newspaper ads (and in addition to the intercept), there is no relationship between Radio and Sales;
3. H_0 for "Newspaper": **in the presence of** TV and Radio ads (and in addition to the intercept), there is no relationship between Newspaper and Sales;
4. H_0 for the intercept: **in the absence of** TV, Radio and Newspaper ads, Sales are zero;

versus the 4 corresponding alternative hypotheses:

H_a: There is some relationship between TV/Radio/Newspaper and Sales, or Sales are non-zero in the absence of the other variables.

https://botlnec.github.io/islp/sols/chapter3/exercise1/

Predicting numeric values with Regression

3. Suppose we have a data set with five predictors, $X_1 = \text{GPA}$, $X_2 = \text{IQ}$, $X_3 = \text{Gender}$ (1 for Female and 0 for Male), $X_4 = \text{Interaction between GPA and IQ}$, and $X_5 = \text{Interaction between GPA and Gender}$. The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50, \hat{\beta}_1 = 20, \hat{\beta}_2 = 0.07, \hat{\beta}_3 = 35, \hat{\beta}_4 = 0.01, \hat{\beta}_5 = -10$.

   (a) Which answer is correct, and why?

      i. For a fixed value of IQ and GPA, males earn more on average than females.

      ii. For a fixed value of IQ and GPA, females earn more on average than males.

      iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

      iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

   (b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

   (c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Predicting numeric values with Regression

Exercise 3.3

(a)

$$Y = \beta_0 + \beta_1 \times GPA + \beta_2 \times IQ + \beta_3 \times Gender + \beta_4 \times GPA \times IQ + \beta_5 \times GPA \times Gender$$

For a fixed value of GPA and IQ, the difference between female and male is given by:

$$Y_{female} - Y_{male} = \beta_3 + \beta_5 \times GPA = 35 - 10GPA,$$

which depends on GPA. It is clear that in the normal range of the GPA (0 to 4.0), the difference in expected salary between female and male ranges linearly from 35 to -5. In particular, if GPA > 3.5, males earn on average more than females. Therefore, **the correct answer is (iii)**.

https://botlnec.github.io/islp/sols/chapter3/exercise3/

Predicting numeric values with Regression

8. This question involves the use of simple linear regression on the `Auto` data set.

   (a) Use the `lm()` function to perform a simple linear regression with `mpg` as the response and `horsepower` as the predictor. Use the `summary()` function to print the results. Comment on the output. For example:

   i. Is there a relationship between the predictor and the response?
   ii. How strong is the relationship between the predictor and the response?
   iii. Is the relationship between the predictor and the response positive or negative?
   iv. What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95 % confidence and prediction intervals?

   (b) Plot the response and the predictor. Use the `abline()` function to display the least squares regression line.

   (c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

Predicting numeric values with Regression

Exercise 3.8

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf  # Statsmodels is a Python module for statistics
import statsmodels.api as sm

%matplotlib inline
```

```
df = pd.read_csv('../data/auto.csv') # import dataset
```

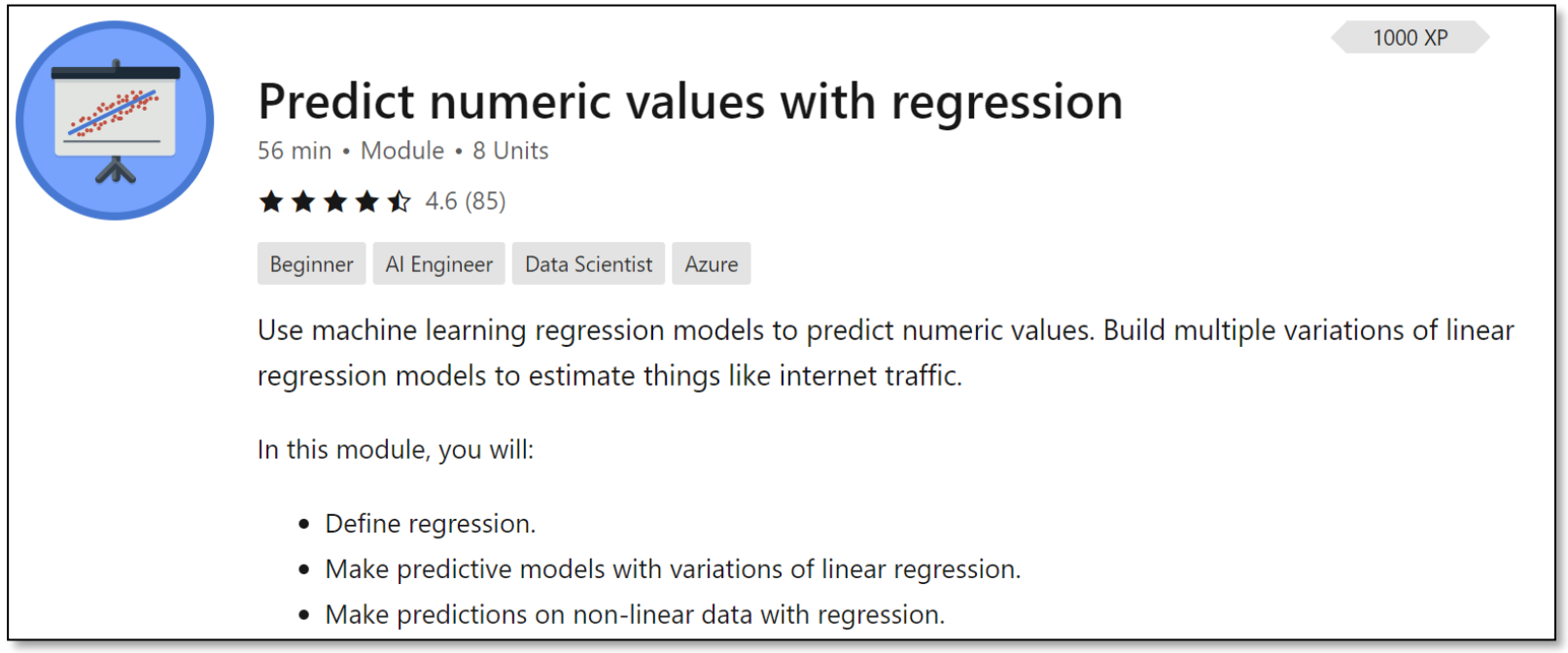
```
df.head() # just to have a look
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino

https://botlnec.github.io/islp/sols/chapter3/exercise8/
http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf | Chapter 3

Exercise 3

Predicting numeric values with Regression



Review of Regression

Introduction

100 XP

5 minutes

Regression can be thought of as looking for a trend line in our data. When used in machine learning, that line allows us to predict a numeric value based on features. We could use regression to predict things like:

- How much money something will cost
- What's the expected temperature tomorrow
- How many days until you need to replace your car battery

For example, if we had a data set with just two features, year and temperature, we could plot those features on a chart and create a trend line that would help us predict the temperature of future years.

Review of Regression

Predict values with simple linear regression

200 XP

6 minutes

The most fundamental type of regression is linear regression. Linear regression tries to draw a straight line as near to the data points as possible.

What is simple linear regression

Simple linear regression is when we use one feature (x) to predict y. Below we have a data set with men's height (x) and shoe size (y):

Exercise for Linear Regression

Exercise - Simple linear regression

100 XP

8 minutes

We'll try out simple linear regression in Azure Notebooks - predicting customer satisfaction with different chocolate bar recipes.

The exercises for this module use Azure Notebooks. It is a free-to-use resource.

If the links below don't work, you will need to set up your free Azure Notebooks account and clone the project for this module. See the Python GitHub repository or the R GitHub repository for instructions to get started.

Review of Multiple Linear Regression

Predict values with multiple linear regression

200 XP

5 minutes

Linear regression using multiple features is called **multiple linear regression**. Multiple linear regression is similar to simple linear regression, but rather than just using one feature to predict a label, it uses multiple features.

Consider a scenario where you work for a government regulatory agency. You're responsible for identifying diesel vehicles with highly toxic emissions. You want to create a model that will predict which vehicles are likely to produce more emissions. Those vehicles will be subject to expensive emissions testing.

A basic version of this predictive model could use two features, weight and horsepower to make the prediction. The graph below shows that we actually fit a plane instead of a line when working with two input features. We can think of a plane as a two-dimensional extrusion of a line, so we still think of this as "linear" regression.

Exercise for Multiple Linear Regression

Exercise - Multiple linear regression

100 XP

10 minutes

Multiple linear regression is commonly used in science, is quick to run, and many modern tools can take care of the hard parts for you.

The exercises for this module use Azure Notebooks. It is a free-to-use resource.

If the links below don't work, you will need to set up your free Azure Notebooks account and clone the project for this module. See the Python GitHub repository or the R GitHub repository for instructions to get started.

Review of non-linear regression

Make predictions on non-linear data using polynomial regression

100 XP

6 minutes

Linear regression is easy to use, but more complex data sets may require more complex algorithms. Many real life data sets require a non-linear model. For example, in the plot below you can see that a straight line doesn't accurately predict coffee sales.

Exercise for non-linear regression

Exercise - Polynomial regression

100 XP

12 minutes

The exercises for this module use Azure Notebooks. It is a free-to-use resource.

If the links below don't work, you will need to set up your free Azure Notebooks account and clone the project for this module. See the Python GitHub repository or the R GitHub repository for instructions to get started.