Exercises and solutions: *Least-squares*

The only way to learn mathematics is *to solve math problems*. Watching and re-watching video lectures is important and helpful, but it's not enough.

Below are some practice problems to solve. You can find many more by searching the Internet.

Exercises

Imagine you work as a data analyst in a company that sells widgets online. The company gives you a dataset of information from 1000 sales, which includes the time of the sale (listed in hours of the day using a 24-hour clock, e.g., 15 = 3pm), the age of the buyer (in years), and the number of widgets sold. The data are included in the "widget_data.mat" file.

1. Explain and write down a mathematical model that is appropriate to test with this dataset.

2. Write the matrix equation corresponding to the model, and describe the columns in the design matrix.

3. Compute the model coefficients using the least-squares algorithm in MATLAB or Python. You can also divide the β coefficients by the standard deviations of the corresponding independent variables, which puts the various model terms in the same scale and therefore more comparable.

4. Produce scatter plots that visualize the relationship between the independent and dependent variables.

5. One measure of how well the model fits the data is called R^2 ("R-squared"), and can be interpreted as the proportion of variance in the dependent variable that is explained by the design matrix. Thus, an R^2 of 1 indicates a perfect fit between the model. The definition of R^2 is
$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$
e_i is the error (residual) at each data point — the difference between the data and the model-predicted data, and $\bar{y}$ is the average of all elements in y.

Compute R^2 for the model to see how well it fits the data. Are you surprised, based on what you see in the scatterplots?

Answers

1. A simple model would predict that time and age are both linear predictors of widget purchases. There needs to be an intercept term, because the average number of widgets purchased is greater than zero.
The variables could interact (e.g., older people buy widgets in the morning while younger people buy widgets in the evening), but I won't include that here in the interest of brevity.

The model would look like this:
$y = \beta_1 + \beta_2\mathbf{t} + \beta_3\mathbf{a}$
y is the number of widgets sold, β_1 is the intercept, $\mathbf{t}$ is the time of day, and $\mathbf{a}$ is the age.

2. The matrix equation is $\mathbf{X}\beta = \mathbf{y}$. $\mathbf{X}$ has three columns: intercept, which is all 1's; time of day; age.

3. MATLAB codes is below. Note that you have to transpose the data matrix.

```
% load the data
load('widget_data.mat')

% design matrix
X = [ones(1000,1) data(1:2,:)'];

% outcome variable
y = data(3,:)';

% beta coefficients
beta = (X'*X) \ (X'*y);
% scaled coefficients (note: scaled intercept is not interpretable)
betaScaled = beta'./std(X);
```

4. MATLAB code below, and see Figure 1 after the answers.

```
subplot(121)
plot(X(:,2),y,'o','markerfacecolor','k')
axis square, title('Time variable')
xlabel('Time of day'), ylabel('Widgets purchased')

subplot(122)
plot(X(:,3),y,'o','markerfacecolor','k')
axis square, title('Age variable')
xlabel('Age'), ylabel('Widgets purchased')
```

5. MATLAB code is below. The model accounts for 36.6% of the variance of the data. That seems plausible given the variability in the data that can be seen in the graphs (Figure 1).

```
yHat = X*beta;
r2 = 1 - sum((yHat-y).^2) / sum((y-mean(y)).^2);
```

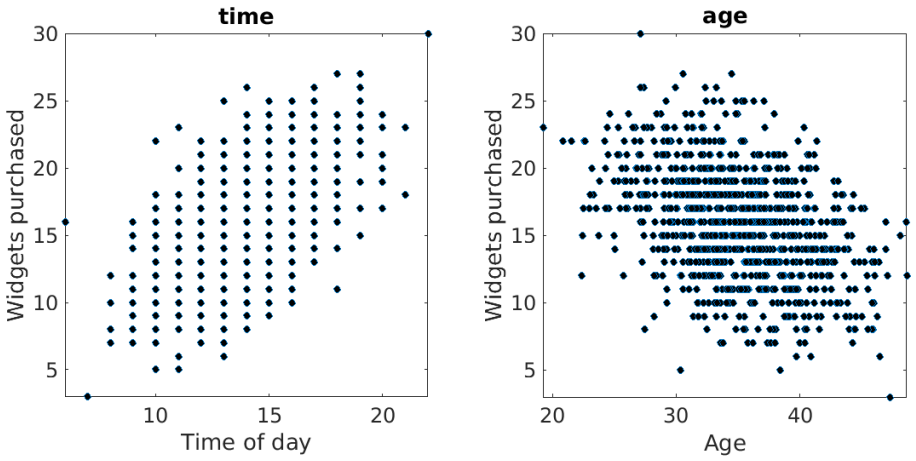

Figure 1: Image for question 4