Linear Regression and Projection Techniques

In this lab assignment, we will cover simple linear regressions and projection of data using common transformations. By the end of the lab you should

1. Be reasonably comfortable with running a simple linear regression model on data as well as assessing the results to see whether the predicted fit is good or not.
2. Be able to transform the popular projection models into a linear form, fit them to data, and make projections.

This is the exact same process you will need to go through when you make projections for your own region.

Regression

First, we will need to download a function performs regression from the class webpage. On the wepbage is file called 'regres.m'. Download it to your computer. You can save it anywhere, but I will give an example where I assume that is is saved in the `C:\` directory.
A little on how matlab works:

When you give matlab a name such as **log**, it doesn't know exactly what to do with it. The first thing matlab does is check to see if it is a variable in the workspace. If there is no variable in the workspace, it then checks a list of directories, called the "path," to see if the function is there. If it isn't, there either, then Matlab returns an error. This is also why we need to make sure not to give a variable the same name as a function we want to use.

Now that we downloaded this function called **regres**, we need to tell matlab where it is. To do this, write:

```
>> addpath 'C:\'
```

If you didn't save it to `C:\`, then place the proper folder between the single quotation marks.

With this done, you should be ready to go. Lets first see how regres works. Write **help regres** You should see a little description on what the program does, how to use, and what it returns.

Lets test it on an exact linear relation.

```
>> X = 0:1:20;
>> X = X';
>> Y = 10+2*X;
>> plot(X,Y);
>> result = regres(Y,X)
>> result.b
>> result.R2
>> result.CI95
```

regres returns a special type of variable, called a structure. A structure is a collection of other objects, such as vectors or matrices. To see what objects are stored in regres, write **¿¿ regres**. The first object, **results.b** is a **2x1** vector. To actually see this vector, write **regres.b**. Note that we need the structure name **regres**, followed by a dot and the variable name **b**.

Note from the plot, that this is an exact linear relation. What are the estimated coefficients? Compare them with the true coefficients of 10 and 2. What is the R^2 value? What are the 95% confidence intervals for the coefficients? You can see that if the relationship is exact, linear regression identifies the line perfectly.

Now, add some normally distributed random error to the relation.

```
>> e1 = randn(21,1);
>> Y1 = 10+2*X+e1;
```

Make a histogram of the error

```
>> hist(e1)
>> plot(X,Y1)
>> result = regres(Y1,X)
```

What are the estimated coefficients? Are they the same as the true values? What is the R^2 value? How big are the 95% confidence intervals?

Now, increase the amount of noise. We will now add 3 times as much noise as we did before.

```
>> e2 = 3 * randn(21,1);
>> Y2 = 10+2*X+e2;
>> plot(X,Y2);
>> result = regres(Y2,X)
```

Compare these results with the previous results. Are we more or less confident in our results when there is more noise?

Transformations and Projections

In this section we will simulate some population data, and then attempt to identify the true model. We will the estimate the parameters of this model and project the model into the future.

DATA SIMULATION:
Write:

```
>> t = [0:0.2:40]';
>> e = (1/20)*randn(201,1);
>> P = log(190)+log(0.9)*t + e;
>> P = 200-exp(P);
>> plot(t,P)
```

How does the data look? Does it appear linear? Exponential? Does it seem to have a ceiling? What model do you think would be most appropriate?

Geometric Model

The functional form for the geometric model is:

$$P = b_0 * b_1^t \tag{1}$$

Or, after taking the log of both sides:

$$\ln(P) = \ln(b_0) + \ln(b_1) * t \tag{2}$$

Turning this into the language of regression, we let $y = \ln(P), b = \ln(b_0), m = \ln(b_1)$ and $x = t$, to ge the formula for a line, $y = b + m * x$

INPUT EVALUATION:
The geometric model assumes that the growth rate is constant, i.e. that:

$$z = \frac{P(t+1) - P(t)}{P(t)} \tag{3}$$

We will see if this holds for our model, write:

```
>> diff = (P(2:201)-P(1:200))./P(1:200);
>> x = [1:200]';
>> plot(x,diff)
```

Does the growth rate appear to be constant?

ORDINARY LEAST SQUARES REGRESSION:
Now, lets fit the model. Write:

```
>> y = log(P);
>> results = regres(y,t);
```

What is the R^2? What are the OLS (i.e. the regression) coefficients? What are the coefficients of the geometric model? To get these, we must transform the OLS coefficients: $b_0^{geom} = \exp(b_0^{OLS}), b_1^{geom} = \exp(b_1^{OLS})$.

PROJECTION:
Now project the transformed model to time = 60, and retransform the dependent variable in order to get the projected population.

```
>> that = [0:0.2:60]';
>> yhat = results.b(1)+results.b(2)*that;
>> Phat = exp(yhat);
>> plot(that,Phat)
```

And plot the original data over the projection

```
>> hold on
>> plot(t,P);
```

Does the projected data fit the original data? Would you believe this projection?

Gompertz Model

The Gompertz model is the first model that places a ceiling on population growth. We will let the ceiling be represented by **c**. In practice, we will beed to pick a reasonable ceiling ahead of time. The functional form for the Gompertz model is:

$$P = c * b_0^{b_1^t} \tag{4}$$

After taking the log of both sides, we get:

$$\ln(P) = \ln(c) + \ln(b_0) * b_1^t \tag{5}$$

After subtracting log(c) from both sides and dividing by (-1), we get:

$$\ln(c) - \ln(P) = \ln(b_0) * (b_1^t) \tag{6}$$

and after taking the log of both sides again, we have a linear equation:

$$\ln(\ln(c) - \ln(P)) = \ln(\ln(b_0)) + \ln(b_1) * t \tag{7}$$

Turning this into the language of regression, we let $y = \ln(\ln(c) - \ln(P)), b = \ln(\ln(b_0)), m = \ln(b_1)$, and $x = t$ to get the linear equation $y = b + m * x$.

INPUT EVALUATION:
The Gompertz model assumes that there is a constant growth rate for the increments, i.e.

$$z = \frac{\ln(P(t+1)) - \ln(P(t))}{\ln(P(t)) - \ln(P(t-1))}$$

Does this hold for our data? Write:

```
>> top = log(P(3:201)) - log(P(2:200));
>> bottom = log(P(2:200)) - log(P(1:199));
>> z = top./bottom;
>> x = [1:199]';
>> plot(x,z);
```

Does this rate appear constant?

OLS REGRESSION:
Before you calculate this model, you will first need to find a good ceiling for the projection. You can come up with a good guess by looking at the plot of the population. From the OLS coefficients, calculate the coefficients of the Gompertz model. Assess the OLS fit. What is the R^2?. Project the linear trend to t = 60. Transform this into population. (Hint, $P = \exp(\log(c) - \exp(y))$.) Does the projection fit the data well? Does it look like a reasonable projection?

Repeat this process with the following models.

Geometric with constant

For the geometric with constant model, the formula is:

$$P = c - b_0 * b_1^t \tag{8}$$
$$c - P = b_0 * b_1^t \tag{9}$$
$$\ln(c - P) = \ln(b_0) + \ln(b_1) * t \tag{10}$$

How do we transform this into a regression (i.e., what should y and x be?).

INPUT EVALUATION:
The Geometric with constant model assumes that the ration of growth increments is constant, i.e.

$$z = \frac{P(t+1) - P(t)}{P(t) - P(t-1)}$$

Does this hold for our data? Write:

```
>> top = P(3:201)-P(2:200);
>> bottom = P(2:200)-P(1:199);
>> z = top./bottom;
>> x = [1:199]';
>> plot(x,z);
```

OLS REGRESSION and PROJECTION:
Determine the proper transformation for this model. Assess its fit. Project the population to time = 60 and assess its fit as well.

Logistic

For the logistic model, the formula is:

$$P = \frac{1}{\frac{1}{c} + b_0 * b_1^t} \tag{11}$$
$$\frac{1}{P} = \frac{1}{c} + b_0 * b_1^t \tag{12}$$
$$\frac{1}{P} - \frac{1}{c} = b_0 * b_1^t \tag{13}$$
$$\ln\left(\frac{1}{P} - \frac{1}{c}\right) = \ln(b_0) + \ln(b_1) * t \tag{14}$$

How do we transform this into a regression (i.e., what should y and x be?).

INPUT EVALUATION:
The logistic model assumes that the ratio of increments of 1/P is constant, i.e.

$$\frac{P(t+1)^{-1} - P(t)^{-1}}{P(t)^{-1} - P(t-1)^{-1}}$$

Does this hold for our population data?
OLS REGRESSION and PROJECTION:
Determine the proper transformation for this model. Assess its fit. Project the population to time = 60 and assess its fit as well.

Final Assessment

Which model had the best R^2 value. Does it also give a reasonable projection of the population?