TAMS65 Project

Elliot Magnusson, ellma121@student.liu.se

May 1, 2020

Contents

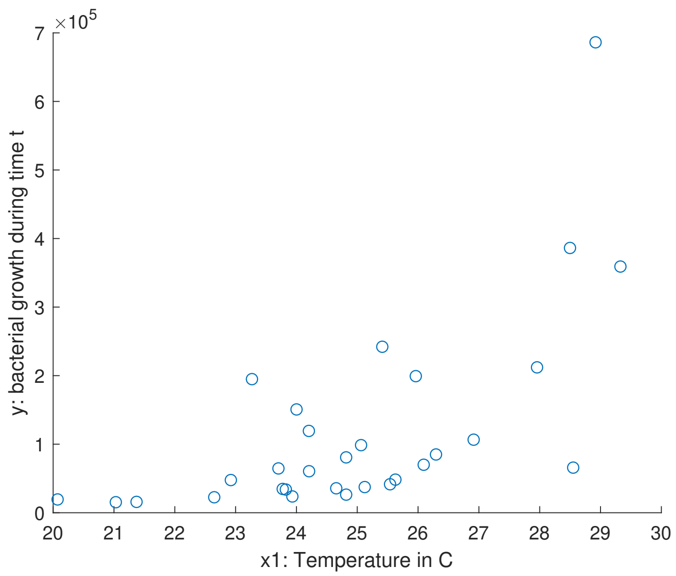
Part 1

Assignment 1: Transformation of Data

In this assignment we are looking how two variables, temperature and environment humidity affects bacterial growth over time.

a)

We first scatter plot the data of x1 against y with `scatter(x1,y)` and get plot below:



We see clear indication of exponential bacterial growth as temperature rises, which leads us to transform y with `y=log(y)` to reveal a nice linear relationship that is nice for further modeling.

Correlation calculated to `0.6459` calculated with `corr(x1,y)`

b)

Using following code:

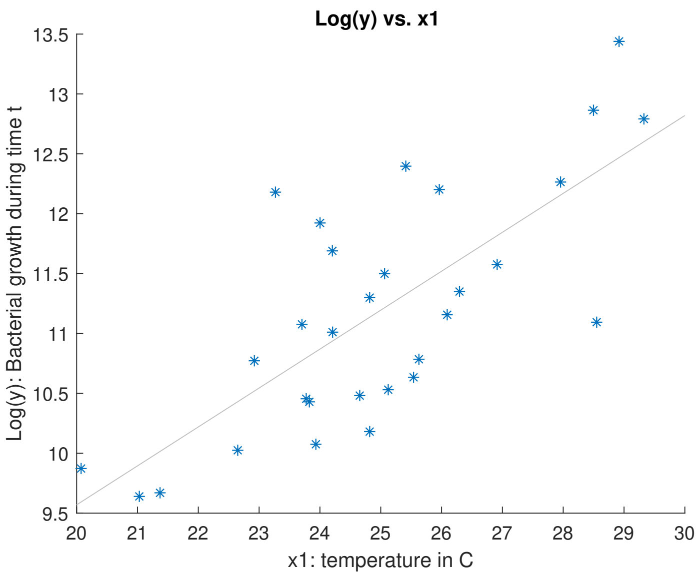
```
tbl = table(logy,x1,x2, 'VariableNames',{'logy','x1','x2'})
tbl.x2 = categorical(tbl.x2) % since binary
mdl = fitlm(tbl, 'logy ¬ x1 + x2')
```

We propose a regression model:

$$log(y) = 1.1720 + 0.3849x_1 + 1.0057x_2 + \epsilon \tag{1}$$

This model has an Rsquared of 0.78, which indicate that proposed regression model explains 78% of variability in response variable $log(y)$. This is good.

c)



Correlation: `0.7404` calculated with `corr(x1,log(y))` and scatterplot with least-squares line shows clear linear relationship.

d)

How many bacteria can we predict for a summer day with the temperature of 25C and low humidity. Calculate an appropriate interval to answer the question

We are seeking a prediction interval for a new observation according to

$$I_{logy} = \boldsymbol{u'}\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u'}(\boldsymbol{X'X})^{-1}\boldsymbol{u}} \tag{2}$$

With $\alpha = 0.05$ and $n - k - 1 = 27$

Our prediction interval then becomes:
$I_{logy} = [9.7739; 11.8168]$ Which re-transformed becomes
$I_y = 1 * e^5[0.1757; 1.3550]$

```
mdl = fitlm(tbl, 'logy¬ x1 + x2')
u = [1 25 0]';
s2 = mdl.MSE;
s = sqrt(s2);
dfe = mdl.DFE
t = tinv(0.975,dfe);
betahat = mdl.Coefficients.Estimate
Cbetahat = mdl.CoefficientCovariance
XtXinv = Cbetahat/s2

%Prediction interval for log(y) = beta0 + beta1*x1 + beta2*x2
I_logy = [u'*betahat-t*s*sqrt(1+u'*XtXinv*u), ...
    u'*betahat+t*s*sqrt(1+u'*XtXinv*u)]
I_Y = exp(I_logy);
```

Assignment 4: Dummy Variables

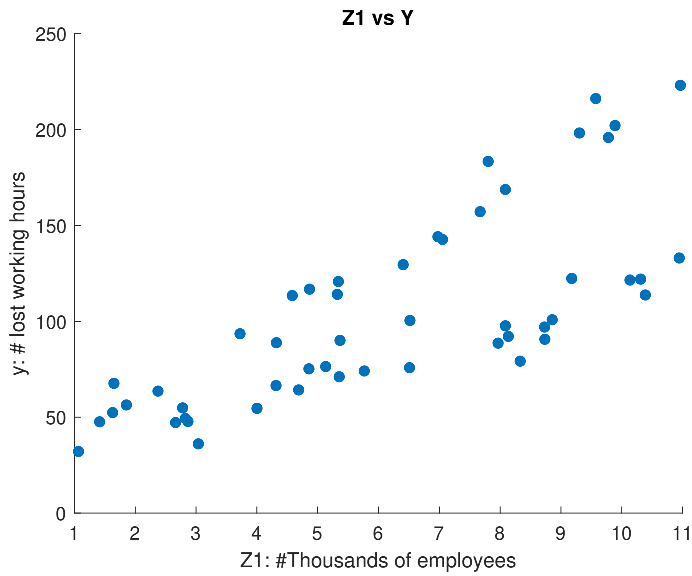
In a study, one wanted to study whether an active security program has significance for the number of working hours which lost due to accidents at work. 50 companies were randomly selected.

a) Analyze according to $Y = \gamma_0 + \gamma_1 z_1 + \gamma_2 x_2 + \epsilon$

With

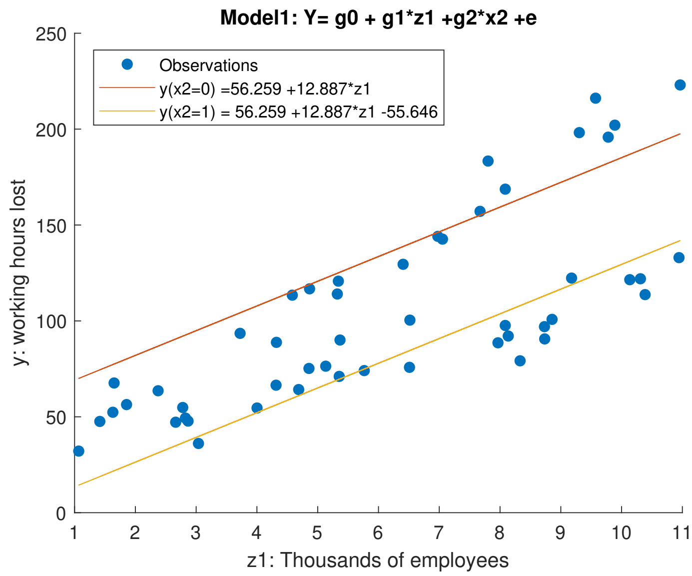
$$z_1 = x_1/1000, \text{and } x_2 = \begin{cases} 1, & \text{if person has security program installed} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Scatter plot y vs z1 clearly shows a linear relationship, along with possible indication of other factor in play to explain variance.



Proposed regression model: $y = 56.251 + 12.887z_1 - 55.646x_2 + \epsilon$ Has $R^2 = 0.89$, which tells us that this model explains 89% of variation in y. I.e Model fits data well.

b) Plot estimated regression lines for $x_2 = 1$ and $x_2 = 0$



c) Analyze according to $Y_2 = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \epsilon$

With

$$z_1 = x_1/1000, \text{and } z_2 = x_2 z_1 \tag{4}$$

Our new regression model 2 becomes as following

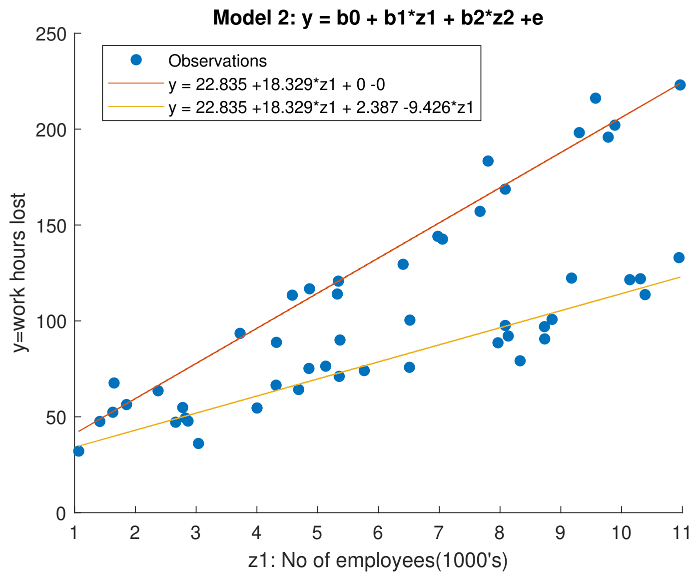
$$y_2 = 22.835 + 18.329 z_1 + 2.387 x_2 - 9.426 z_1 x_2 \tag{5}$$

with

$$\begin{cases} 1, & \text{if person has security program installed} \\ 0, & \text{otherwise} \end{cases}$$

With $R^2 = 0.97$ it seems this model explains the variation in y even better than model 1.

d) Plot regression lines of $x_2 = 1$ and $x_2 = 0$



We see that the line fits data better than in model 1, which is reasonable since r^2 is higher.

e) Does use of software indicate fewer working hours lost?

We construct an confidence interval for β_3

```
%Significance test on b3 since the slope of the lines are the ...
    difference we are looking for.
% H0: b3 = 0 (for the interaction term z1:x2)
% H1: b3 /= 0

>> anova(mdl2)

ans =

  4  5  table

                SumSq     DF     MeanSq       F          pValue
                ------    --     ------     ------     ----------

    z1          67738      1      67738     808.44     7.8432e-31
    x2          37713      1      37713      450.1     2.1569e-25
    z1:x2      8843.9      1     8843.9     105.55     1.7121e-13
    Error      3854.3     46     83.789


% null hypothesis is rejected at the 5% significance level =>
% slopes not equal => security programs seem to reduce working ...
    hours lost.

%Another test with confidence intervals
betaCI=coefCI(mdl2);
betaCI =

   13.3258   32.3434    %beta0
   16.9256   19.7322    %beta1
  -10.1489   14.9233    %beta2
  -11.2731   -7.5794    %beta3
```

CI for b_3 does not include zero, which indicates that using security program does lower working hours lost on a 5% significance level.

Part 2

Assignment 2

a) [None needed]

b)

Q: Give a suitable linear regression model with response variable y1 and explanatory variable x1.

A: $Y_b = 29.188 + 19.598x_1$ $R^2 = 62.7\%$

c)

$Y_c = 131.991 - 86.321x_1 + 25.787x_1^2$
$R^2 = 98.9\%$ Much better than previous model.

d)

Q: Consider all observations. Give a suitable linear regression model and calculate the coefficient of determination R2.

A: $y_d = 91.670 - 23.935x_1 + 479081396991467x_1^2 - 479081396991463x_1^3$

e)

Q: Calculate the stationary points for the strength, and state at intervals what strength we can expect for these currents.

A:

Local stationary points from looking at regression-line plot $x_max = -1$ with $y_max = 116$ $x_min = 1.4$ with $y_min = 67.1$

I.e for currents between x = [-2 ; 0] we can expect strength of 116 For currents between x = [1;2] we can expect strengths of 67.1

Assignment 3

a)

$y = 79.2089 + 1.0631x_1 + 0.5477x_2$ $R^2 = 0.507$

b)

-

c)

$y = 79.8225 + 1.0675x_1 + 0.4161x_2 + 0.7493x_1x_2 - 1.1999x_1^2 - 0.4545x_2^2$ $R^2 = 0.749$

d)

Testing $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$
against H_1 : one of $\beta_3, \beta_4, \beta_5 \neq 0$ on 95% confidence level. We reject w as $= 14.1 > 4.6 = c$

New variables seems useful.

e)

Max yield of 87.0757 for $x_1 = x_2 = 1$

I.e we would choose reaction time of 80, and temperature of 150C.

Assignment 5

a)

Correlation y against $x_i, i = 1, .., 8$
corr = [1.0 0.5479 0.0282 -0.3762 -0.1321 -0.4290 0.6276 0.605 -0.1842]

b)

$R^2 = 0.9978$, Root mean squared error $= 1.67$

c)

$y = 0.2623 + 10.1x_1 + 1.6634x_2 - 6.7139x_4 + 11.7114x_6 + 3.6675x_7 + 0.2831x_8$ $R^2 = 0.9978$

d)

Test: $H0 : \beta_3 = \beta_5 = 0$ $H1$: at least one of$\beta_3, \beta_5 \neq 0$

We reject H0 on 5% level. The full model does not seem to describe the data better.

Assignment 6

a)

Correlation y against $x_i, i = 1, .., 8$
corr = [1.0 0.5479 0.0282 -0.3762 0.1321 -0.4290 0.6276 0.6057 -0.1842]

b)

$R^2 = 0.9978$, Root mean squared error $= 1.67$

c)

$y = 0.2623 + 10.1x_1 + 1.6634x_2 - 6.7139x_4 + 11.7114x_6 + 3.6675x_7 + 0.2831x_8$ $R^2 = 0.9978$

d)

Test: $H0 : \beta_3 = \beta_5 = 0$ $H1 :$ at least one of$\beta_3, \beta_5 \neq 0$

We reject H0 on 5% level. The full model does not seem to describe the data better.

Assignment 7

a)

$R^2 = 0.9978$

b)

$y = -3.746 + 10.217x_1 + 1.661x_2 - 7.161x_4 + 14.977x_5 + 11.283x_6 + 3.794x_7 + 0.304x_8$
$R^2 = 0.998$

c)

Test: $H0 : \beta_3 = 0$ $H1 : \beta_3 \neq 0$

We reject H0 on 5% level($w = 0.3958 < 8.0166 = c$). The full model does not seem to describe the data better than all-subsets regression.