Regression Analysis Project

Maximilian Sam Licke Agdur

Spring 2020

Part I

Problem 5: Forward Selection

For an industrial process the measure of chemical yield has a set of possible explanatory variables. In doing regression analyzes we want to choose the best model, by performing a forward selection algorithm we can achieve this. The definition of the response variable and the definitions of the explanatory variables follow below.

$Y =$ measure of chemical yield,
$x_1 =$ amount of catalyst,
$x_2 = \begin{cases} 1, & \text{preprocessing 2;} \\ 0, & \text{otherwise} \end{cases}$
$x_3 = \begin{cases} 1, & \text{preprocessing 3;} \\ 0, & \text{otherwise} \end{cases}$
$x_4 =$ humidity i %,
$x_5 =$ percentage of oxygen in the surrounding environment,
$x_6 =$ time in seconds for the process,
$x_7 =$ square of time of process,
$x_8 =$ temperature i $^\circ C$.

In order to get a better grasp in what way the response variable depends on each and every possible explanatory variable we perform a scatter plot of y against x_i $\forall i \in \{1, 2, 3, 4, 5, 6, 7, 8\}$ and calculate the correlations between the two. Data for 30 observations have been given to use. Programming in Matlab we use the function corr(.) to find a vector of empirical correlations which we display as a row vector below. Using the function scatter(x,y) we find the scatter plots described above, these are put on page 3.

$$corr = \begin{pmatrix} 0.5479 & 0.0282 & -0.3762 & -0.1321 & -0.4290 & 0.6276 & 0.6057 & -0.1842 \end{pmatrix} \tag{1}$$

Starting with the regression analysis of this process we begin by finding a multiple linear regression model in which all possible explanatory variables are included. This model can be expressed generally as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

Using the code regstats(y, [x1 x2 x3 x4 x5 x6 x7 x8], 'linear', 'all') we find that the best linear model that includes all eight explanatory variables is the following estimated regression hyperplane:

$$\hat{y} = -5.20 + 10.25x_1 + 1.75x_2 + 0.50x_3 - 70.3x_4 + 15.45x_5 + 11.38x_6 + 3.79x_7 + 0.30x_8$$

We call this model the total model. We find that the coefficient of determination is $R^2_{total} = 0.997764$ which is an almost perfect value, i.e. close to 1.0.

A key assumption of the linear regression model is that the residuals are normally distributed and independent, that is to say $e_i = y_i - \hat{y}_i \overset{iid}{\sim} N(0, \sigma)\ \forall i$. Looking at a histogram of the residuals, on page 4, there is no obvious pattern, we could, however, say that they approximately look like they follow some bell curve. This is also justified by looking at the qq-plot of the residuals that show a approximate normal relationship. In addition we look at the residuals plotted against the fitted values there is no pattern which supports the assumption that the residuals have constant variance.

We move on by finding a second model by using forward selection. Summarized the forward selection works in the following way:

1. Start with a given or empty model. Out of the set of possible new explanatory variables, pick the one with the greatest absolute correlation with respect to the response variable. Test this variable by testing $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$. If the test rejects H_a pick the next best correlated variable and repeat.

2. Add one new variable and calculate the SSE of the model for every possible variable left. Then pick the model that yields the smallest SSE. Again, test that the chosen variable is useful. If it was, add it to the model. If it was not pick the next smallest SSE yielding variable.

3. Repeat step 2 until there are no useful explanatory variables left.

Applying the algorithm to the given data set the following model was ascertained. We call this the forward model. Below the general model we also provide the estimated regression hyperplane.

$$Y = \beta_0 + \beta_6 x_6 + \beta_1 x_1 + \beta_2 x_2 + \beta_7 x_7 + \beta_4 x_4 + \beta_8 x_8 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

$$\hat{y} = 0.26 + 11.71 x_6 + 10.09 x_1 + 1.66 x_2 + 3.67 x_7 - 6.71 x_4 + 0.28 x_8$$

The new regression model has a coefficient of determination $R^2_{forward} = 0.997548$. By visual inspection of the residuals, they seem to be normally distributed looking at the histogram and qq-plot of the forward model's residuals on page 4. A visual inspection of the scatter plot of fitted values by the forward model and its residuals shows no obvious pattern which supports the assumption of constant variance.

In order to determine whether or not the forward selected model is better or worse than the total model we will compare the coefficients of determination as well as conducting a hypothesis test on the beta coefficients. We note that $R^2_{total} > R^2_{forward}$ which seems to indicate the total model is better. Since the total model has more explanatory variables than the forward selected model we choose to test: $H_0 : \beta_3 = \beta_5 = 0$ against H_a : at least one $\beta_i \neq 0\ \forall i$. Again we choose the significance level $\alpha = 0.05$.

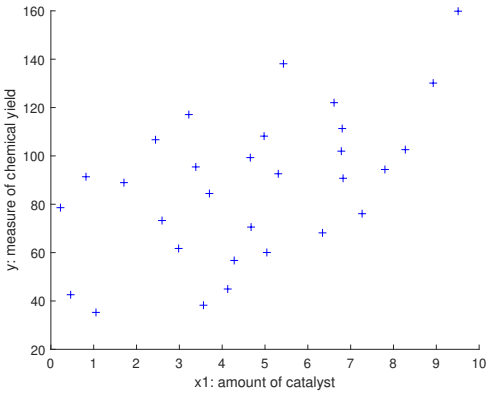
$$Model1(forward) : Y = \beta_0 + \beta_6 x_6 + \beta_1 x_1 + \beta_2 x_2 + \beta_7 x_7 + \beta_4 x_4 + \beta_8 x_8 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

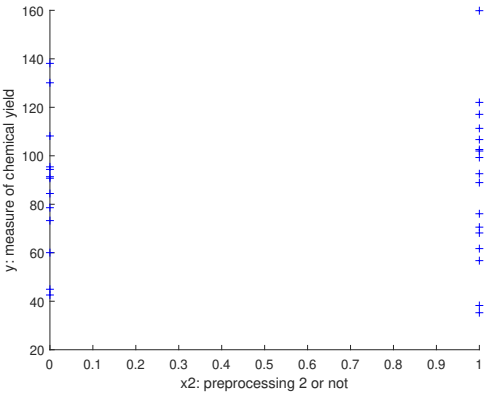
$$Model2(total) : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

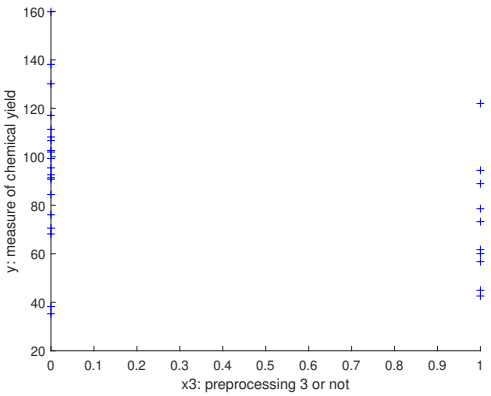
Our test statistic is:

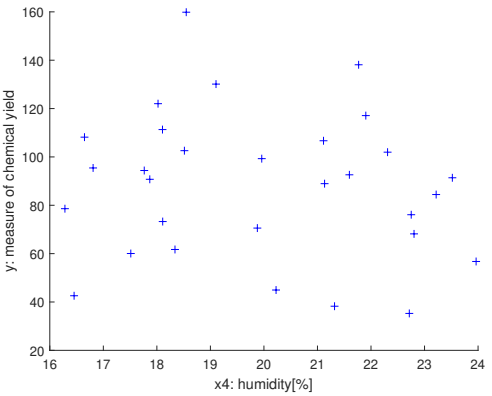
$$W = \frac{(SSE^1 - SSE^2)/p}{SSE^2/(n-k-p-1)} \sim F(p, n-k-p-1)$$

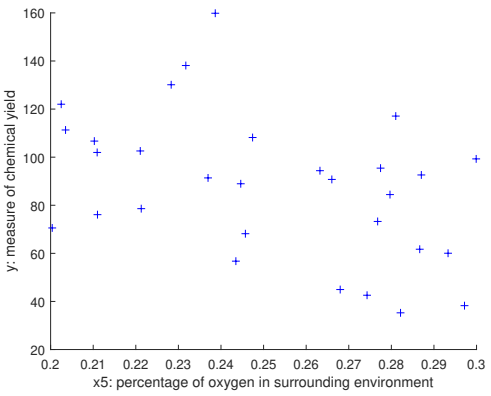
in which SSE^1 is the SSE from model 1 and SSE^2 is the SSE from model 2, p is the number of additional variables, n is the number of observations, k is the number of explanatory variables in model 1. Since $n = 30$, $k = 6$, $p = 2$, $SSE^1 = 64.255$ and $SSE^2 = 58.586$ we get $W = 1.0159$. The critical region is given by $C = (F_\alpha(p, n-k-p-1), \infty) = (F_{0.05}(2, 21), \infty) = (3.47, \infty)$. Note that $W \notin C$, therefore we cannot reject $H_0 : \beta_3 = \beta_5 = 0$ which means that the modified model is not significantly better than the total model, even though the coefficient of determination of model 1 is a bit larger than the coefficient of determination of model 2.

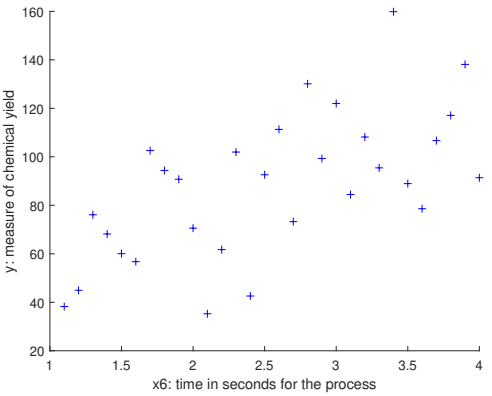

y: measure of chemical yield
x1: amount of catalyst

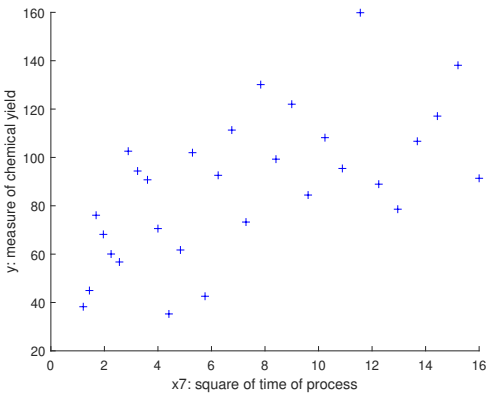

y: measure of chemical yield
x2: preprocessing 2 or not

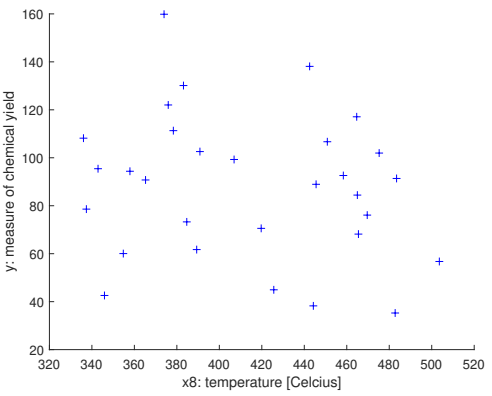

y: measure of chemical yield
x3: preprocessing 3 or not

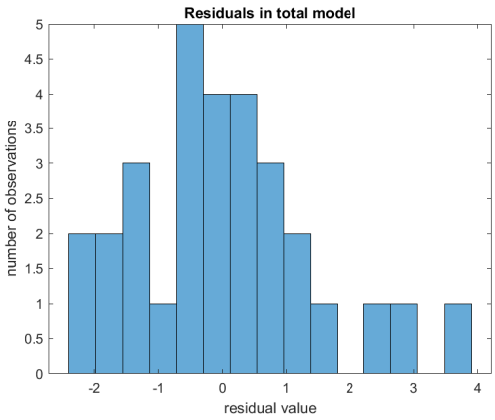

y: measure of chemical yield
x4: humidity[%]

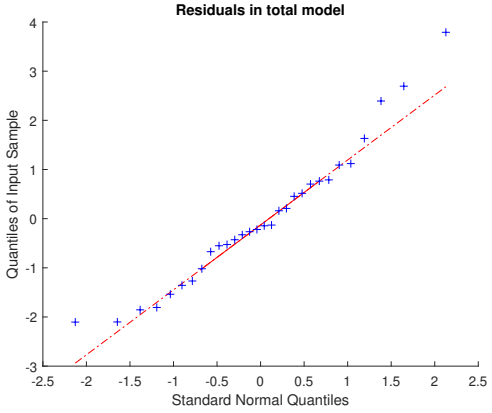

y: measure of chemical yield
x5: percentage of oxygen in surrounding environment

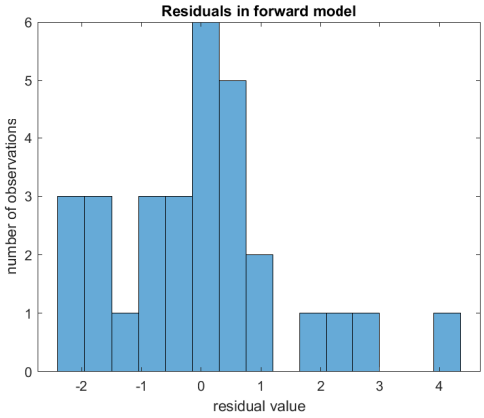

y: measure of chemical yield
x6: time in seconds for the process

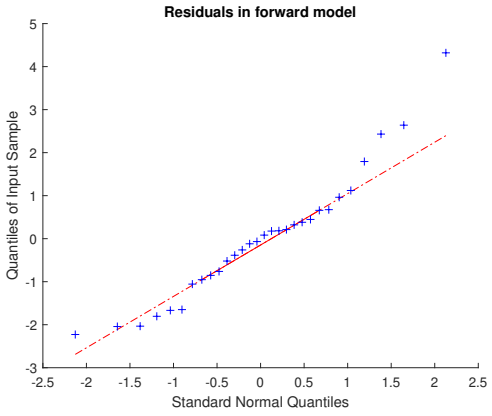

y: measure of chemical yield
x7: square of time of process

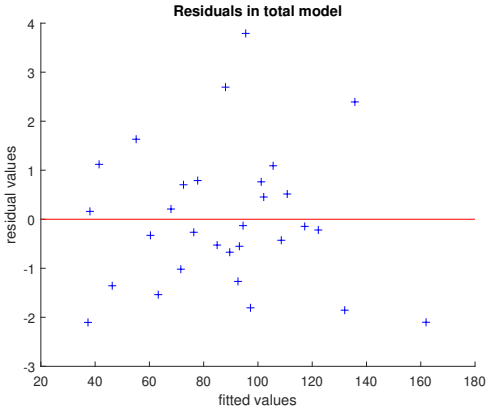

y: measure of chemical yield
x8: temperature [Celcius]

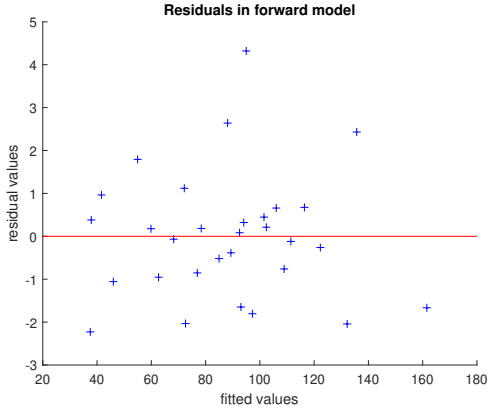

Residuals in total model
number of observations
residual value


Residuals in total model
Quantiles of Input Sample
Standard Normal Quantiles


Residuals in forward model
number of observations
residual value


Residuals in forward model
Quantiles of Input Sample
Standard Normal Quantiles


Residuals in total model
residual values
fitted values

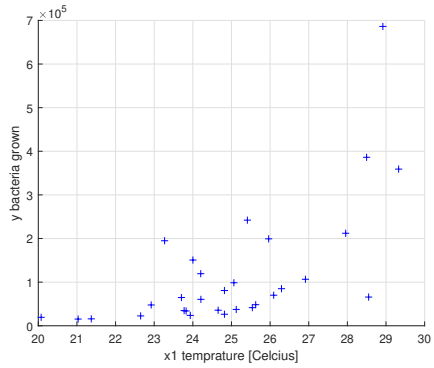

Residuals in forward model
residual values
fitted values

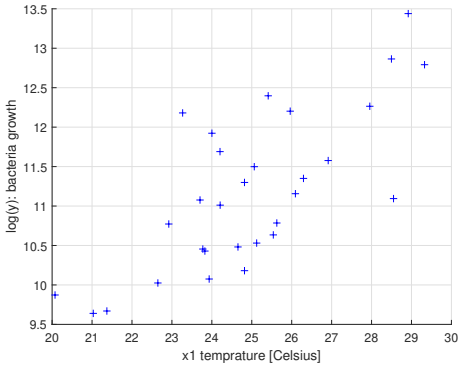
Problem 1: Transformation of Data

We are analyzing under which conditions bacteria has a favorable growth. We are given 30 independent observations and the following definitions of the response variable and the explanatory variables.

$$
\begin{aligned}
y &= \text{the number of bacteria grown during a fixed time} \\
x_1 &= \text{temperature i}^{\circ}C, \\
x_2 &= \begin{cases} 0 & \text{too low humidity} < 80\%, \\ 1 & \text{for high humidity} \geq 80\%. \end{cases}
\end{aligned}
$$

We begin by examining the correlations between the variables. Firstly using a scatter plot of y and x_1 we find the bottom left graph in which there seems to be some positively correlated behaviour. Notably we see more data points clustered around the bottom from 20 to 26 degrees Celsius, and the rest grow quite quickly in a seemingly exponential manner. The positive correlation is accentuated by finding the empirical correlation $corr = 0.6459$ with the Matlab function corr(x1,y).





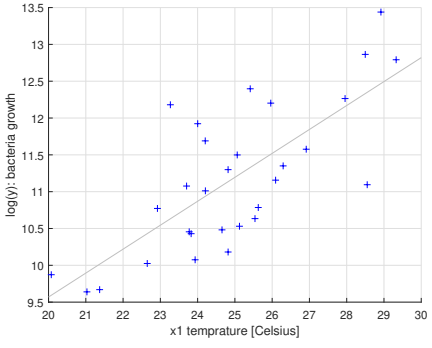
Using the transformation $ln(y)$ and then preforming the same correlation analysis as with the non-transformed variable we observe that the correlation is more linear. This is supported by the empirical correlation being higher at $corr = 0.7404$ suggesting that the exponential increase was true. Performing a linear regression analysis on this new transformed data set with x1 and x2 as explanatory variables with the model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \varepsilon \sim N(0, \sigma)$$

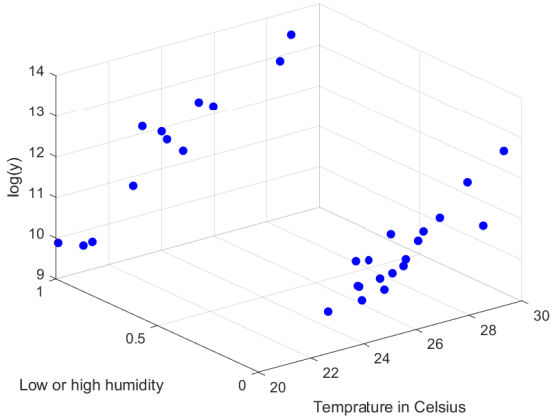
Using the regstats(log(y), [x1 x2], 'linear', 'all') function we end up with the following estimated regression plane:

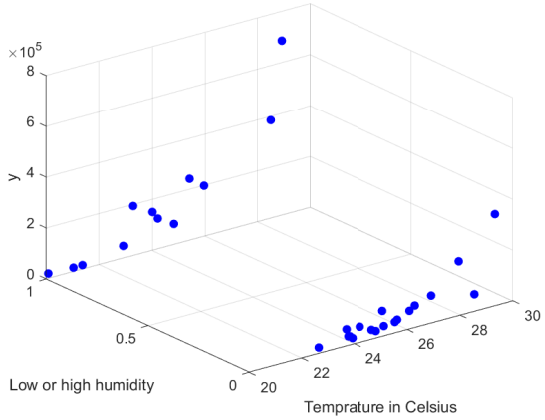
$$\hat{y} = 1.1720 + 0.3849 x_1 + 1.0057 x_2$$

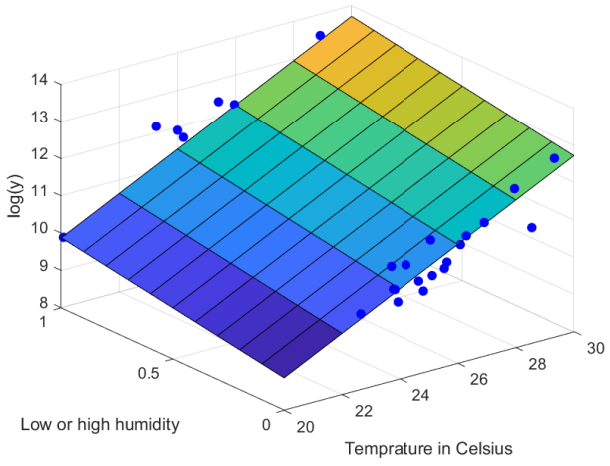
Having a coefficient of determination $R^2 = 0.7812$. We may now plot the estimated regression plane with respect to x_1 which is done below. We can observe a feasible estimated regression line.



Before moving on with predicting values we make some further observations. Displayed below is a scatter plot of all data points with respect to the transformed y value, the normal y value ; as well as the estimated regression plane. One notable thing is that irrespective of high or low humidity the growth in bacteria seems to grow in a similar exponential manner which can be seen on the right figure.







Now we are going to predict the number of bacteria after the experimental time (i.e. the same time as all other observations) given low humidity and a temperature of 25 degrees Celsius. Note that both $x_1 = 25$ and $x_2 = 0$ satisfy $x_i^{min} \leq x_i \leq x_i^{max}$, i.e. they are in the range of the values of the observations. Writing this as an observation vector we find $\boldsymbol{u^T} = \begin{pmatrix} 1 & 25 & 0 \end{pmatrix}$. Recall that the prediction interval of a new observation is given by the following formula:

$$I_{log(Y_0)}^{1-\alpha} = \boldsymbol{u}^T\boldsymbol{\beta} \pm t_{\alpha/2}(n-k-1)s\sqrt{\boldsymbol{u}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{u}+1}$$

In which $\boldsymbol{X}$ is the design matrix. With 30 observations $n = 30$, two explanatory variables $k = 2$ and a significance level of $\alpha = 0.05$ we see that $t_{0.025}(27) = 2.0518$. s can be estimated with $s^2 = \frac{SSE}{n-k-1}$ which gives us $s = 0.4842$. Using Matlab to complete the rest of the calculation we end up with the following prediction interval:

$$I_{log(Y_0)}^{0.95} = (9.7739, 11.8167)$$

However, since we transformed the data we need to reverse that to find the actual number of bacteria. The prediction interval above is the log of the number of bacteria. Using $exp()$ we can find the prediction interval of the actual number of bacteria.

$$I_{Y_0}^{0.95} = (17569, 134592)$$

Note that the prediction interval may vary greatly depending on the number of decimals used in the interval. This can be seen since $exp(11.8167) = 134592$ but $exp(11) = 59874$ and so on.

Part II

Problem 2: Polynomial Regression

1. The model is $Y = \beta_0 + \beta_1 x_1 + \varepsilon, \varepsilon \sim N(0, \sigma)$. The estimated regression line is: $\hat{y}_1 = 29.19 + 32.41 x_1$
2. Use a quadratic polynomial. The model is $Y_{quad} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon, \varepsilon \sim N(0, \sigma)$. The estimated regression polynomial is: $\hat{y}_{quad} = 131.99 - 86.32 x_1 + 25.79 x_1^2$. $R^2_{quad} = 0.9890$
3. Use a cubic polynomial. The model is $Y_{cube} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon, \varepsilon \sim N(0, \sigma)$ The estimated regression polynomial is: $\hat{y}_{cube} = 99.32 - 26.31 x - 3.37 x^2 + 4.11 x^3$. $R^2_{cube} = 0.9872$
4. Stationary points for the cubic estimated regression polynomial are: $x = 1.76$ and $x = -1.21$. corresponding to the currents of $z = 6759.9$ and $z = 3786.6$ respectively. Prediction intervals for these points are: $I_{Y_0} = (55.49, 74.46)$ and $I_{Y_0} = (109.43, 128.46)$ respectively.

Problem 3: Response Surface Regression

1. The model is: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \varepsilon \sim N(0, \sigma)$ The estimated regression plane is: $\hat{y} = 79.21 + 1.06 x_1 + 0.55 x_2$. $R^2 = 0.51$.
2. The model is : $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon, \varepsilon \sim N(0, \sigma)$ The estimated regression surface is: $\hat{y} = 79.82 + 1.07 x_1 + 0.42 x_2 - 1.20 x_1^2 - 0.45 x_2^2 + 0.75 x_1 x_2$. $R^2 = 0.75$.
3. We test $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$ for $i = 3, 4, 5$ for $\alpha = 0.05$. Our critical region is $C = (-\infty, -2.0154) \cup (2.0154, \infty)$. Finding the test statistics for each and every test above we can conclude that x_1^2 and $x_1 x_2$ are useful while x_2^2 is not useful.
4. Using constrained optimization (quadratic programming) with $\hat{y} = 79.82 + 1.07 x_1 + 0.42 x_2 - 1.20 x_1^2 - 0.45 x_2^2 + 0.75 x_1 x_2$ subject to $-1 \leq x_i \leq 1 \quad \forall i$ we obtain a global maximum at $(1, 1)$ corresponding to $t_1 = 100[seconds]$ and $t_2 = 150[Celsius]$.

Problem 4: Dummy Variables

1. $R^2 = 0.89$.
2. $R^2 = 0.97$.
3. $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$ with $\alpha = 0.05$. We know $I^{0.95}_{\beta_2} = (-9.26, -8.96)$ since $0 \notin I^{0.95}_{\beta_2}$ we discard H_0. It looks like $\beta_2 < 0$ this is a indication that the security programs are leading to fewer working hours lost.

Problem 6: Backward Elimination

1. corr $= \begin{pmatrix} 0.5479 & 0.0282 & -0.3762 & -0.1321 & -0.4290 & 0.6276 & 0.6057 & -0.1842 \end{pmatrix}$
2. $R^2 = 0.997764$
3. The model obtained was: $Y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \beta_6 x_6 + \beta 7 x_7 + \beta_8 x_8 + \varepsilon, \varepsilon \sim N(0, \sigma)$. The estimated regression hyperplane became: $\hat{y} = -2.75 + 10.25 x_1 - 7.85 x_4 + 12.25 x_6 + 3.57 x_7 + 0.34 x_8$ $R^2 = 0.997202$
4. Testing $H_0 : \beta_2 = \beta_3 = \beta_5 = 0$ against H_a : at least one $\beta_i \neq 0$. Our critical region with $\alpha = 0.05$ is $C = (3.15, \infty)$ and our test statistic is $T = 1.51$. Since $T \notin C$ we cannot reject H_0. We cannot conclude that the full model is significantly better.

Problem 7: All Subsets Regression

1. $R^2 = 0.9978$
2. The model obtained was: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon, \varepsilon \sim N(0, \sigma)$. The estimated regression hyperplane became: $\hat{y} = -3.75 + 10.22x_1 + 1.66x_2 - 7.16x_4 + 14.98x_5 + 11.28x_6 + 3.79x_7 + 0.30x_8$
3. Testing $H_0 : \beta_3 = 0$ against $H_a : \beta_3 \neq 0$. With significance level $\alpha = 0.05$ Our critical region is: $C = (4.3248, \infty)$. But our test statistic is $T = 0.3958$, thus cannot reject H_0, and the total model isn't significantly better than the model from all possible subsets.