Data Analysis for Mechanical Engineering

Regression Concepts

William 'Ike' Eisenhauer

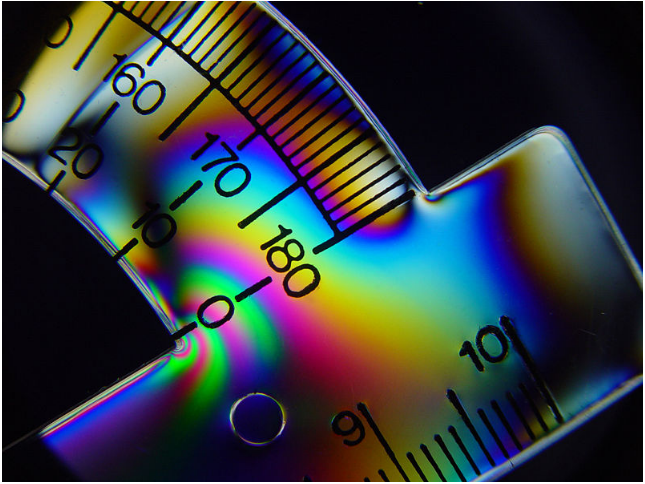
Department of Mechanical and Materials Engineering
Portland State University
Portland, Oregon 97223

wde@pdx.edu

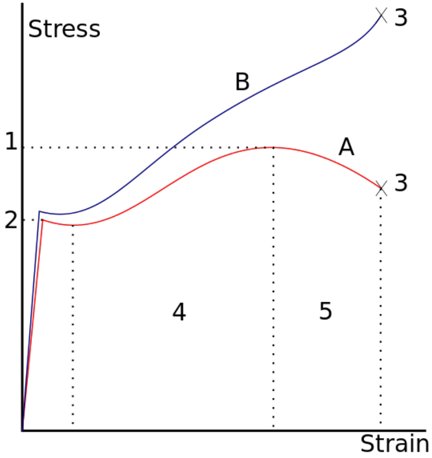
Winter 2016

Bourdon Tube Gauge

Protractor



Stress vs Strain



Whats in the Tank?

An unknown gas, of known mass, is in a thin walled pressure vessel of constant volume. Determine what's in the tank without opening it and potentially killing everyone.

$$PV = nR_u T$$

$$PV = nM_{gas}RT$$

$$PV = m_{gas}RT$$

$$P = \frac{m_{gas}R}{V}T$$

$$P = (Constant \cdot R)T$$

Key Aims

We want to:

- Understand **linear** regression with one predictor
- Understand how we assess the fit of a regression model
  - Total Sum of Squares
  - Model Sum of Squares
  - Residual Sum of Squares
  - F
  - r^2
- Interpret a Regression Model Output

Classic Viewpoint

Method of Prediction

A way of predicting the value of one variable from another, using a hypothetical model of the relationship between two variables.

Keep It Simple Stupid!

The basic models we will use assume the relationship is linear, and thus we can describe the relationship as an equation of a line.

Engineering Viewpoint

Functional Relationship

In engineering, we use it to also establish the functional relationship, using a hypothetical model of the relationship.

Key focus is the Slope

The slope for our purposes helps define a relationship.

Basic Linear Model

Form of the Basic Linear Model

$$Y_{model} = \beta_1 X_1 + \beta_0 + \varepsilon$$

β_1: Regression Coefficient for the Predictor

Gradient (slope) of the regression line [direction and strength of the relationship].

β_0: Intercept

For many engineering purposes we assume this is 0 [Origin Assumption], unless we have content validity that it may not be.

ε: Error of the model (Residual)

Very rarely is the model going to match the data perfectly

What is a good model?

Well, anyone can slap a line on a graph and call it a model. And even if all you have is some measure of central tendency, you can make a really basic model.

Basic Mean Model

$$Y = \bar{Y} + \varepsilon$$

This is considered the most naive model that all others are to be compared. And we should be able to do better.

Variability

So now we have three things we can measure:

Total Variability

The variability between the data and the mean. Total variance in the data.

Residual/Error Variability

The variability between the data and the model. Error in the Model

Model Variability

The variability between the model and the mean. Improvement Due to the Model.

So what?

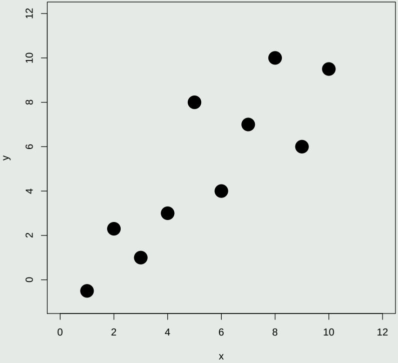
Well, if the model we pick is a better predictor [i.e. describer of the real relationship] than the lame mean model, then we should expect:

$$Variability_{Model} \ggg Variability_{Residual}$$

In other words, the Improvement the Model gives is better than the error in that Model, or forget it!

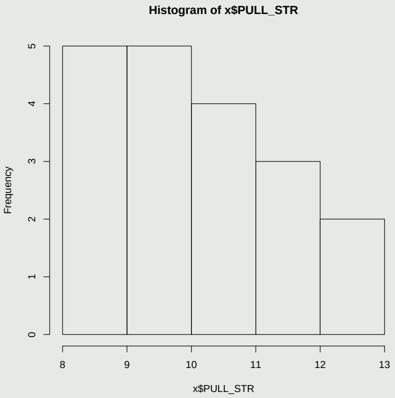
Visualize the Concept

The Data



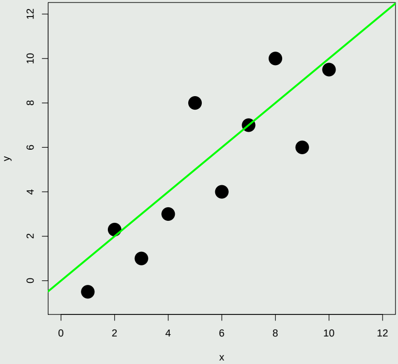
Visualize the Concept

The Mean



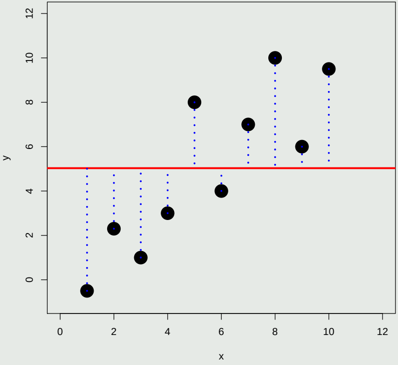
The Model

The Model



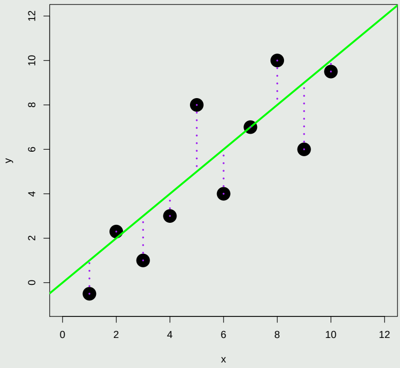
SS_T: Variability between Data and Mean

SS_T: Total Variance in Data



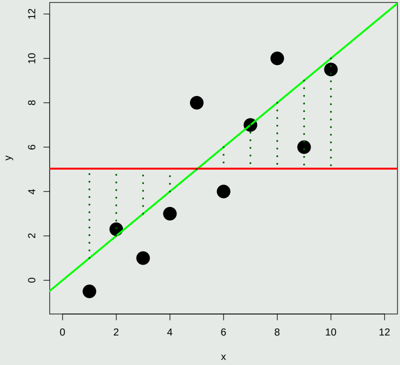
SS_R: Variability between Data and Model

SS_R: Error in Model



SS_M: Variability between Model and the Mean

SS_M: Improvement in Model [over Mean]



F: Model test by ANOVA

Mean Squared Error

The Sums of Squares were TOTAL values, if we divide them by the degrees of freedom we have they are called **Mean Squares**: MS_M and MS_R

Ratio of MS_M and MS_R

We hope that our Model will have high improvement over 'mean model' and low error in describing the data.

$$\frac{MS_M}{MS_R} \ggg 1$$

F: Model test by ANOVA

ANalysis Of VARiance: ANOVA

These are variances, use ANOVA F-test to see if it really is bigger than 1.

$$H_0 : F = \frac{MS_M}{MS_R} = 1$$

$$H_A : F = \frac{MS_M}{MS_R} \ggg 1$$

F-test: Checks if the Model is Meaningful

Remember, F is used to see if the model you come up with is any better than the naive mean model. In other words, is it meaningful. There is a highly unprofessional, but effective way to remember this...

r^2: Model test by Coefficient of Determination

Usefulness

Once you have determined that the model is meaningful. You might want to know if it is **useful**. We need to compare the variance captured by the model, that is in the data to begin with.

Ratio of SS_M and SS_T

We hope that our Model variance will be very close to the Total variance

$$\frac{SS_M}{SS_T} = 1$$

r^2: Model test by Coefficient of Determination

Deja Vu?

So where have we seen something like that before?

$$r^2 = \frac{SS_M}{SS_T}$$

r^2: Model test by Coefficient of Determination

Deja Vu?

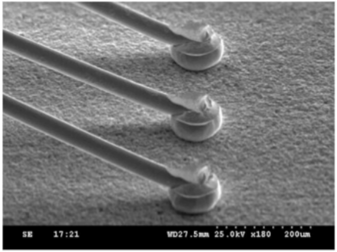
So where have we seen something like that before?

$$r^2 = \frac{SS_M}{SS_T}$$

r^2: The Proportion of Variance Accounted for by the Model

The closer this is to 1 the more plausible it is that the model is describing the data [from a variance perspective]. Be careful if this shows up as $= 1$...bad mojo

Semiconductor Assembly [Palomar]



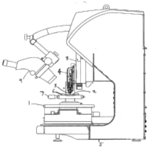


Figure 2

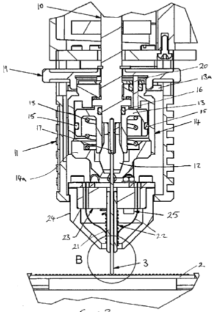


Figure 3

Semiconductor Assembly (Palomar)

Output Variable

Pull Strength

Input Variables

1. Die Height
2. Post Height
3. Loop Height
4. Wire Length
5. Bond D
6. Bond P

Key Aims

We want to:

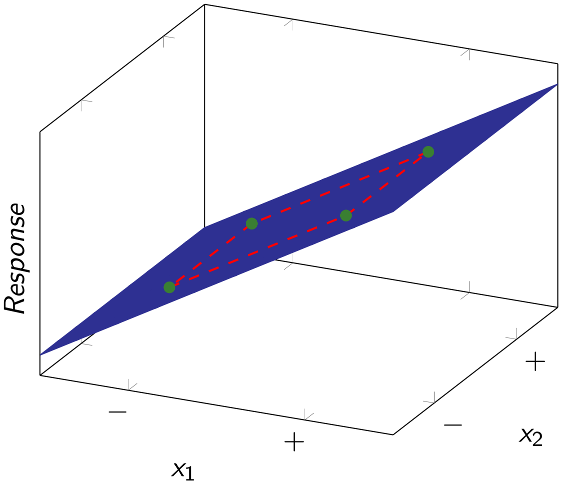
- Understand **linear** regression with multiple predictors
- Understand regression equation and meaning of the βs
- Understand different methods of adding predictors
  - Hierarchical
  - Stepwise
  - Forced Entry

Multiple Linear Regression

Extension of the Single Model

Used to predict values of an outcome from several predictors, using a hypothetical model of the relationship between several variables.

Visualization of Concept



Multiple Linear Model

Form of the Model

$$Y_{model} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

β_i: Regression Coefficient for X_i

Partial Derivative of the regression hyperplane with respect to $Predictor_i$.
A unit change in X_i results in a change of β_i in Y.

Methods of Adding Variables

- Hierarchical
  - Experimenter decides the order in which the variables are entered in the model
- Forced Entry
  - All predictors are entered simultaneously
- Stepwise - BE CAREFUL!
  - Predictors are selected solely on their semi-partial correlation with the outcome

Hierarchical

Hierarchical

- Known predictors (based on past research) are entered into the regression model first
- New predictors are then entered in a separate step/block
- Experimenter makes the decisions

Forced Entry

Forced Entry

- All variables are entered into the model simultaneously.
- Results obtained depend on the variables entered into the model.
- It is important, therefore, to have good theoretical reasons for including a particular variable.

Goals

Meaningful Model

Is equation (model) better than using the mean?
Check $F \ggg 1$ [Rule of Thumb: $F > 8$]

Useful Model

Does equation (model) actually describe the reality?
Check r^2 [Rule of Thumb $adjr^2 > 0.85$]

Model Complexity

Have relatively insignificant variables?
Check $\beta_i \neq 0$

Checking β

If $\beta = 0$?

Then the variable most likely doesn't have much to do with the situation

Confidence Intervals of β

Regression analysis gives the confidence intervals around β. So we are looking at removal candidates where 0 is inside this interval.

P-values of β

Regression analysis also gives you a different view called the **p-value**, but they are basically telling you the same thing, so I personally find it easier to deal with the confidence intervals.

Ditching X_i

If $\beta_i = 0$?

Then the variable most like doesn't have much to do with the situation, so "look" at getting rid of it

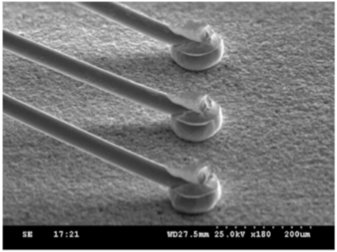
Removing one affects the whole situation

Don't just go all medieval on the variables. Start with most likely to be zero, take it out and run regression again, since the allocations of variance have to shift around

When do I stop?

Stop when removing when your model either becomes meaningless [F too low] or non-useful [r^2 too low].

Semiconductor Assembly [Palomar]



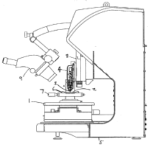


Figure 2

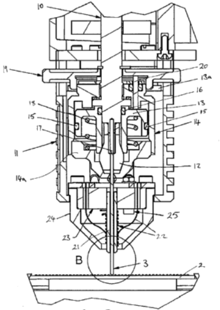


Figure 3

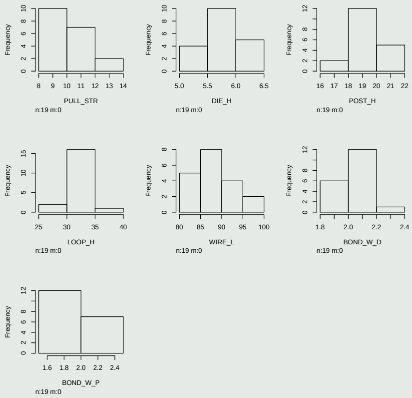
Load and PTDD

R Code

```
library(Hmisc) #Package to make better histograms
mydata <- read.csv("MLR_EXAMPLE_DATA.csv")
hist(mydata)
plot(mydata)
```

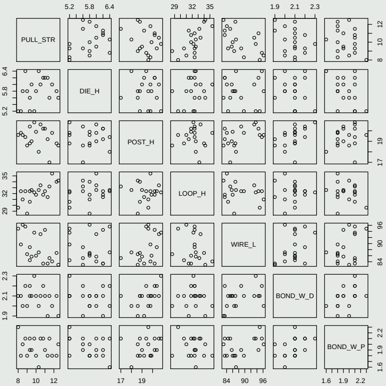
Load and PTDD

Results - Histograms



Load and PTDD

Results - Scatter Matrix



Linear Regression - LM function

R's **lm** function

lm means linear model. It is R's way of doing linear regression. Two key parts.

1. Establish the model (i.e. store it in an R variable)
2. Review the model (i.e. use that variable), via **summary**

Linear Regression - LM function

Establish the model

- Y: output variable
- X_i: input variables
- dataframe: the dataframe your data is in

R syntax

```
model <- lm(Y~X1+X2+....+XN, data=dataframe)
model <- lm(Y~X1+X2+....+XN+0, data=dataframe)
```

Linear Regression - LM function

Engineering Example

```
m <- lm(PULL_STR~DIE_H+POST_H+LOOP_H+WIRE_L
+BOND_W_D+BOND_W_P, data=mydata)
```

Note, no immediate output!!!

Linear Regression - Results

Now get the model output

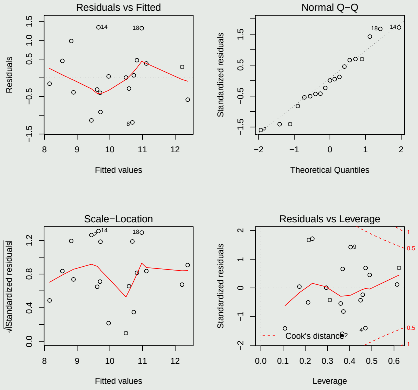
```
summary(m)
par(mfrow=c(2,2))
plot(m)
anova(m)
confint(m)
confint(m, level= 0.90)
```

Linear Regression - Summary

Results - Summary

```
##
## Call:
## lm(formula = PULL_STR ~ DIE_H + POST_H + LOOP_H + WIRE_L + BOND_W_D +
##     BOND_W_P, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1904 -0.3939  0.0072  0.4180  1.3472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1368     8.1098    0.39    0.706
## DIE_H         0.6444     0.5889    1.09    0.295
## POST_H       -0.0104     0.2677   -0.04    0.970
## LOOP_H        0.5046     0.1423    3.55    0.004 **
## WIRE_L       -0.1197     0.0562   -2.13    0.055 .
## BOND_W_D     -2.4618     2.5978   -0.95    0.362
## BOND_W_P      1.5044     1.5194    0.99    0.342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.894 on 12 degrees of freedom
## Multiple R-squared:  0.711,^^IAdjusted R-squared:  0.567
## F-statistic: 4.93 on 6 and 12 DF,  p-value: 0.00921
```

Results - Plots



Linear Regression - ANOVA

Results - ANOVA

```
## Analysis of Variance Table
##
## Response: PULL_STR
##           Df Sum Sq Mean Sq F value Pr(>F)
## DIE_H      1   3.47    3.47    4.34 0.0593 .
## POST_H     1   1.75    1.75    2.19 0.1650
## LOOP_H     1  13.48   13.48   16.86 0.0015 **
## WIRE_L     1   3.77    3.77    4.71 0.0507 .
## BOND_W_D   1   0.38    0.38    0.48 0.5033
## BOND_W_P   1   0.78    0.78    0.98 0.3416
## Residuals 12   9.59    0.80
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Regression - CI

Results - CI

```
##                   2.5 %    97.5 %
## (Intercept) -14.5329 20.80653
## DIE_H         -0.6387  1.92757
## POST_H        -0.5936  0.57274
## LOOP_H         0.1945  0.81477
## WIRE_L        -0.2422  0.00286
## BOND_W_D      -8.1218  3.19826
## BOND_W_P      -1.8060  4.81482
```

Linear Regression - CI

Results - CI Specific

```
##                  5 %      95 %
## (Intercept) -11.3172 17.59080
## DIE_H        -0.4052  1.69405
## POST_H       -0.4874  0.46661
## LOOP_H        0.2510  0.75833
## WIRE_L       -0.2199 -0.01944
## BOND_W_D     -7.0917  2.16819
## BOND_W_P     -1.2035  4.21236
```

Next time

A few more things we need to do for a real full analysis. And that is to simplify our model [Note the F before], and to use it to predict a value based on data NOT in the original dataset.