STA 674

Regression Analysis And Design Of Experiments

Fitting Multiple Linear Regression Models – Lecture 1

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Where does it fit in?
- What is it?
- Where next?

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

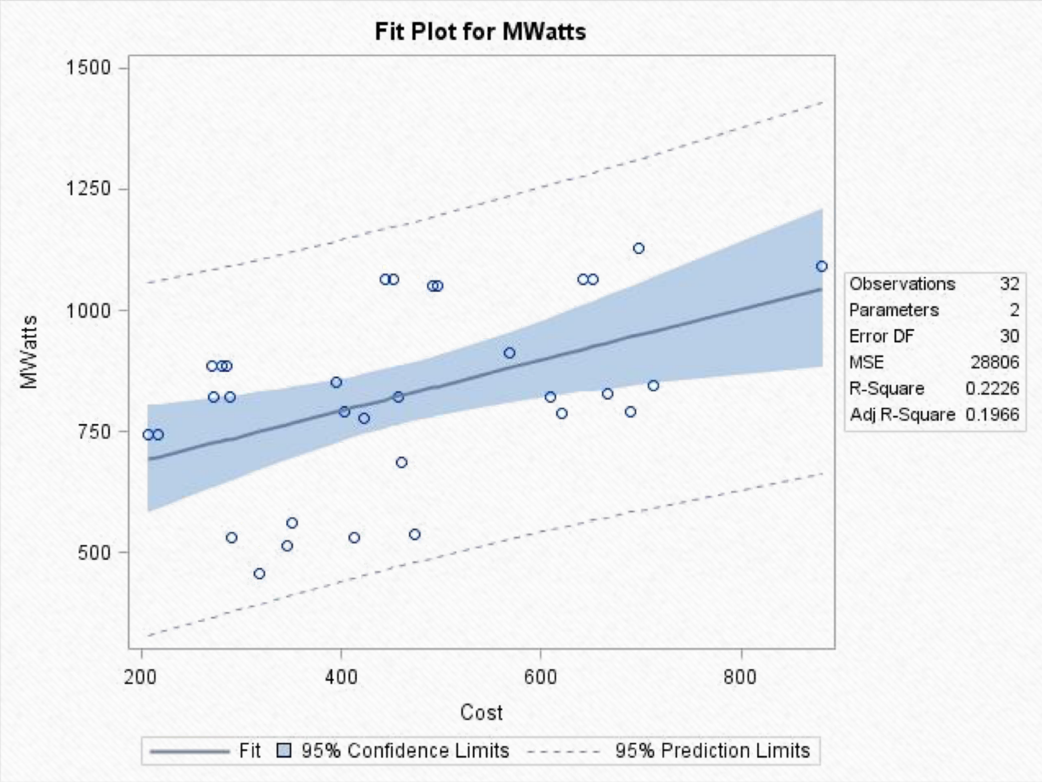
- We will consider a data set containing information on 32 nuclear power plants built between 1967 and 1971. The data contain three variables:
  - **Response:** Output in mega watts (MWatts)
  - **Predictor 1:** Cost in $100,000 (adjusted to 1976) (Cost)
  - **Predictor 2:** Date of construction in years after 1900 (Date)
- Goal: Examine how output varies with cost and year of construction.

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

- SLR – Output versus Cost

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	583.00457	87.97883	6.63	<.0001
Cost	1	0.52511	0.17919	2.93	0.0064

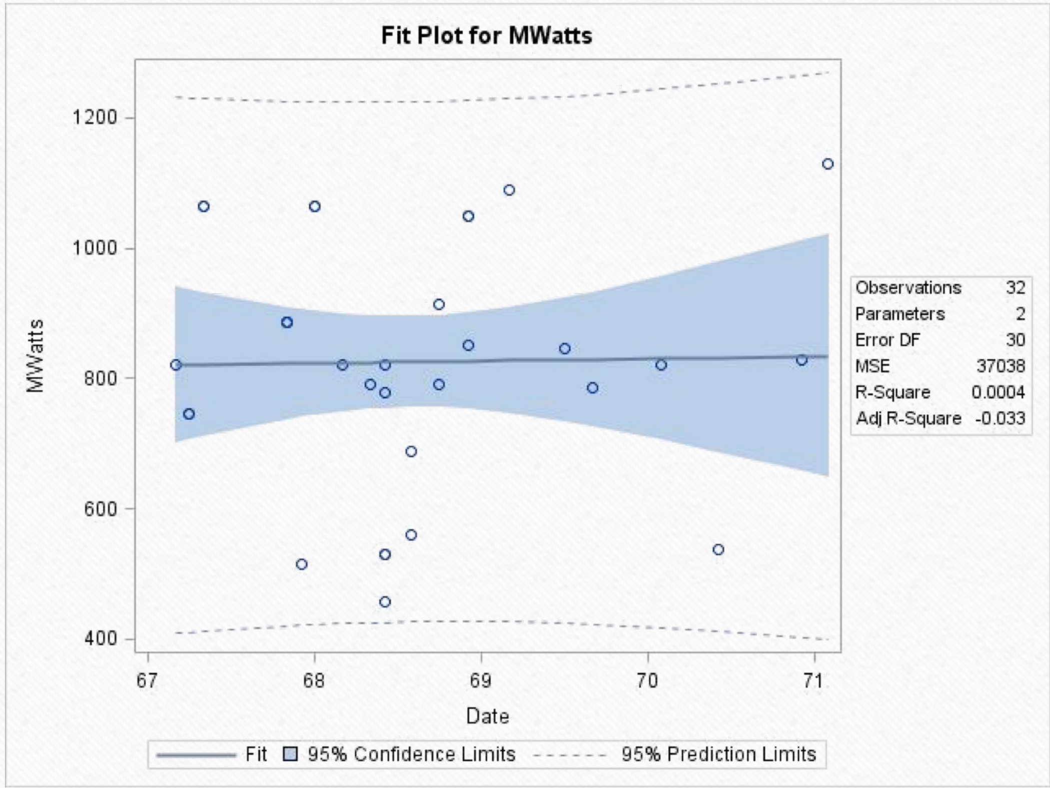


STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

- SLR – Output versus Date

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	574.29406	2335.14259	0.25	0.8074
Date	1	3.66107	34.04567	0.11	0.9151



STA 674, RADOE:
Fitting Multiple Linear Regression Models

Model

- The **multiple linear regression model with *K* predictors** is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \epsilon_i$$

Notation

- y_i - response observed for individual i
- $x_{1i}, x_{2i}, \dots, x_{Ki}$ - predictors for individual i
- β_0 intercept (for all individuals)
- β_j - regression coefficient of $x_j, (j = 1, \dots, K)$
- ϵ_i - residual error for individual i

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Least squares estimation uses the values of $\beta_0, \beta_1, \dots, \beta_K$, called $b_0, b_1, \dots, b_K$, which minimize the sum of the squared errors:

$$SSE = \sum_{i=1}^{n} (\underbrace{y_i}_{\text{observed}} - \underbrace{\hat{y}_i}_{\text{predicted}})^2$$

over all possible values of $b_0, b_1, \dots, b_K$.

Why?

1. Least squares estimates are unique.
2. Least squares estimates can be computed by hand.
3. Least squares estimates are optimal (if the assumptions are satisfied).

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

- MLR – Output versus Cost and Date

```
/* Multiple linear regression model */;
TITLE "3. Mwatts vs Cost and Date";
PROC REG DATA=NUCLEAR;
    MODEL mwatts=cost date / CLB;
RUN;
```

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	5920.53408	2382.75823	2.48	0.0190	1047.24632	10794
Cost	1	0.81578	0.21243	3.84	0.0006	0.38131	1.25025
Date	1	-79.78403	35.59545	-2.24	0.0328	-152.58490	-6.98316

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

- SLR – Output versus Cost
- SLR – Output versus Date
- MLR – Output versus Cost and Date

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	583.00457	87.97883	6.63	<.0001
Cost	1	0.52511	0.17919	2.93	0.0064

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	574.29406	2335.14259	0.25	0.8074
Date	1	3.66107	34.04567	0.11	0.9151

For each passing year, we predict the average output DECREASES by ~80 Mw while holding Cost (predictor) constant

We estimate output increases by 0.8 Mw per $100K increase in cost for a given year (holding other variable...ie, Cost)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	5920.53408	2382.75823	2.48	0.0190	1047.24632	10794
Cost	1	0.81578	0.21243	3.84	0.0006	0.38131	1.25025
Date	1	-79.78403	35.59545	-2.24	0.0328	-152.58490	-6.98316

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Interpreting Parameter Estimates

- The parameter β_j measures the change in y per unit change in x_j after adjusting for the other predictors in the model.
- By this we mean that β_j represents the average change in y per unit change in x_j ***while*** all of other predictors remain the same.
- The value b_j is simply the estimate of this parameter. It is the average change in y that one would expect to see per unit increase in x_j if all other predictors stay fixed.

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

Exercise:

- Provide an interpretation of the least squares estimates for the multiple regression model of output vs cost and date of construction.
- Can you explain why date of construction is a significant predictor in the multiple linear regression model, but not in the simple linear regression model?

Consider: cost versus date