STA 674

Regression Analysis And Design Of Experiments

Fitting Multiple Linear Regression Models – Lecture 2

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Last time, introduced MLR
- Now, we do inference.

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Inference: as with SLR, we start with confidence intervals

Interval Estimate of the Regression Parameter β_k Pick 1 parameter at a time

- A $(1-\alpha)100\%$ confidence interval for β_k has endpoints:

$$L = b_k - s_{b_k} t_{\alpha/2, n-K-1} \text{ and } U = b_k + s_{b_k} t_{\alpha/2, n-K-1}$$

equation for CI is same as SLR...but degrees of freedom is different - n-k-1

where s_{b_k} is the standard error of b_k.

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;
```

Cost:
We are 95% confident that the true slope of the regression model is between 0.351 to 1.25...average output increases between 0.351 and 1.25 for each additional $100K spent (while holding Data fixed)

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 Simple Linear Regression Models

Significance test for the population slope—first, upper-tailed test

- **Hypotheses:** $H_0: \beta_k \leq \beta_k^*$ vs. $H_a: \beta_k > \beta_k^*$ H0 is equal or less than or equal...Ha is always greater than
- **Test Statistic:** $t = (b_k - \beta_k^*)/s_{b_k}$
- **Rejection Rule:** Reject H_0 and conclude that $\beta_k > \beta_k^*$ at the α level of significance if $t > t_{\alpha,n-K-1}$. Otherwise, we fail to reject H_0.
- ***P*-value:** Probability that t-distribution with $n - K - 1$ degrees of freedom is greater than TS t. $\text{p} = P(t_{n-K-1} > t)$

STA 674, RADOE:
Fitting Simple Linear Regression Models

Significance test for the population slope—next, lower-tailed test

- **Hypotheses:** $H_0: \beta_k \geq \beta_k^*$ vs. $H_a: \beta_k < \beta_k^*$
- **Test Statistic:** $t = (b_k - \beta_k^*)/s_{b_k}$
- **Rejection Rule:** Reject H_0 and conclude that $\beta_k < \beta_k^*$ at the α level of significance if $t < -t_{\alpha,n-K-1}$. Otherwise, we fail to reject H_0.
- ***P*-value:** Probability that t-distribution with $n - K - 1$ degrees of freedom is less than TS t. $p = P(t_{n-K-1} < t)$

STA 674, RADOE:

Fitting Simple Linear Regression Models

Significance test for the population slope—finally, two-tailed test**

- **Hypotheses**: $H_0: \beta_{\mathrm{k}} = \beta_{\mathrm{k}}^*$ vs. $H_a: \beta_{\mathrm{k}} \neq \beta_{\mathrm{k}}^*$
- **Test Statistic**: $t = (b_{\mathrm{k}} - \beta_{\mathrm{k}}^*)/s_{b_{\mathrm{k}}}$
- **Rejection Rule**: Reject H_0 and conclude that $\beta_{\mathrm{k}} \neq \beta_{\mathrm{k}}^*$ at the α level of significance if $t < -t_{\alpha/2,n-2}$ or $t > t_{\alpha/2,n-2}$. Otherwise, we fail to reject H_0.
- ***P*-value**: Probability that t_{n-K-1} is "farther away from 0" than TS t.

$$p = 2 \times P(t_{n-K-1} > |t|)$$

- **By default, this is the hypothesis SAS tests.

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

H0: B2(cost) = 0, Ha: B2(cost) != 0

p = 0.0328 ... or ... compare t value with test statistic

Reject the null hypothesis

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Interpreting Significance Tests

- Rejecting H_0 implies that there is a statistically significant effect/positive effect/negative effect of x_k on the response after accounting for the effects of the other variables.
- E.g., if we reject $H_0: \beta_k = 0$ then we can conclude that changing x_k while holding all of the other variables fixed changes the mean response.

Warning!

- Failing to reject $H_0: \beta_k = 0$ does NOT mean that there is no effect of x_k on the response.
- Failing to reject $H_0: \beta_k = 0$ implies that changing x_k while holding all of the other variables fixed does not change the mean response.

STA 674, RADOE:

Fitting Multiple Linear Regression Models

Prediction – Fitted Values and CIs for the Mean

- The fitted value for the mean with values of the predictors $x_{m1}, x_{m2}, \ldots, x_{mK}$ is:

$$\hat{y}_m = b_0 + b_1 x_{1m} + b_2 x_{2m} + \cdots + b_k\, x_{Km}$$

- A $(1-\alpha)100\%$ confidence interval for $\hat{y}_m$ has endpoints:

$$\hat{y}_m \pm t_{\alpha/2,n-K-1} s_m$$

where s_m is the standard error of $\hat{y}_m$.

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Prediction – Predicted Values and Prediction Intervals for new observations

- The best prediction of a single new observation with values of the predictors $x_{p1}, x_{p2}, \ldots, x_{pK}$ is:

$$\hat{y}_p = b_0 + b_1 x_{1p} + b_2 x_{2p} + \cdots + b_k\, x_{Kp}$$ Same as for CI

- A $(1-\alpha)100\%$ confidence interval for $\hat{y}_p$ has endpoints:

$$\hat{y}_p \pm t_{\alpha/2, n-K-1} s_p$$ Standard error is bigger...so bigger PI

  where s_p is the standard error of $\hat{y}_p$.

- As in simple linear regression, $s_p = \sqrt{s_m^2 + s_e^2}$.

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Nuclear Power Plant Data

- Suppose that you used the MLR model predict the output of nuclear reactors built in 1972 at a cost of $500 hundred thousand dollars (adjusted to 1976).

```
TITLE "3. Mwatts vs Cost and Date";
PROC REG DATA=NUCLEAR2;
    MODEL mwatts=cost date / CLB;
    OUTPUT OUT=NUCLEARPRED PRED=pred LCL=ucl UCL=lcl LCLM=uclm UCLM=uclm;
RUN;
```

Obs	Cost	MWatts	Date	pred	uclm	uclm2	ucl	lcl
33	500	.	72	583.971	338.066	829.877	175.667	992.276

Mean output (Mw)

- Provide an interpretation of these results.