STAC67: Regression Analysis

Lecture 17

Sohee Kang

Mar. 17, 2021

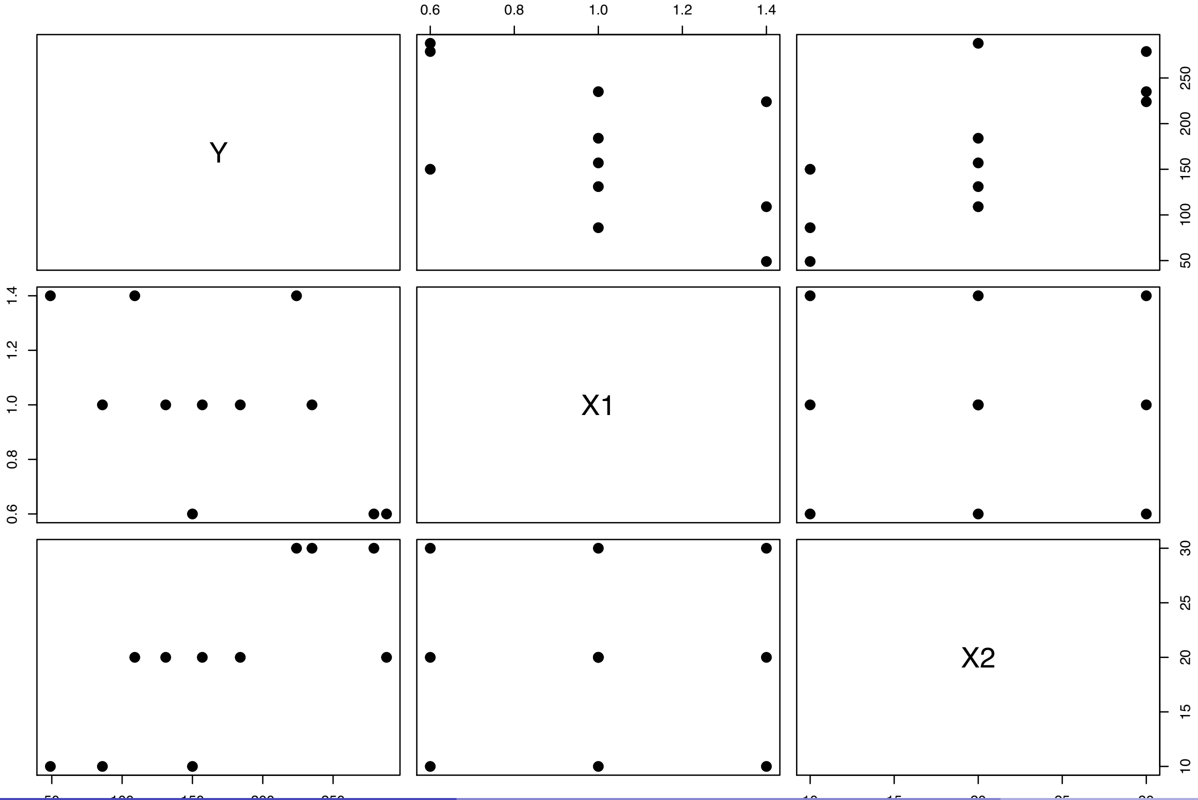
Power cells example

- Researcher studies the effects of the charge rate (amperes) and temperature (degrees Celsius) of a new type of power cell in a preliminary small-scale experiment.
- Three levels of charge rate and of temperature
- Life of the power cell in terms of the number of discharge-charge cycles before the cell failed

Cell i	Number of cycles Y_i	Charge rate X_{i1}	Temperature X_{i2}
1	150	0.6	10
2	86	1.0	10
3	49	1.4	10
$\vdots$	$\vdots$	$\vdots$	$\vdots$
11	224	1.4	30
Mean		1.0	20

Power cells example

```r
Powercell= read.table("Table8-1.txt", header=T)
par(mfrow=c(2,2))
pairs(Powercell, pch=19, cex=1.5)
```



Power cells example

- polynomial second order seems to be a good idea.

```
fit = lm(Y~X1 + X2 + I(X1^2)+I(X2^2)+I(X1*X2), data=Powercell)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2), data = Powercell)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
## -21.465   9.263  12.202  41.930  -5.842 -31.842  21.158 -25.404 -20.465   7.263
##      11
##  13.202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  337.7215   149.9616   2.252   0.0741 .
## X1          -539.5175   268.8603  -2.007   0.1011
## X2             8.9171     9.1825   0.971   0.3761
## I(X1^2)      171.2171   127.1255   1.347   0.2359
## I(X2^2)       -0.1061     0.2034  -0.521   0.6244
## I(X1 * X2)     2.8750     4.0468   0.710   0.5092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.37 on 5 degrees of freedom
## Multiple R-squared:  0.9135, Adjusted R-squared:  0.8271
## F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

None of the predictors are significant in T-test

$H_0: \beta_0 = \cdots \beta_6 = 0$

Model is significant, which is a contradiction in F-test

$\therefore$ There's multicolinearity.

R output

- The correlation matrix of the variables included in the model is:

	Y	X_1	X_2	X_1^2	X_2^2	X_1X_2
Y	1.000	-0.556	0.751	-0.529	0.737	0.255
X_1	-0.556	1.000	0.000	0.991	0.000	0.605
X_2	0.751	0.000	1.000	0.000	0.986	0.757
X_1^2	-0.529	0.991	0.000	1.000	0.006	0.600
X_2^2	0.737	0.000	0.986	0.006	1.000	0.746
X_1X_2	0.255	0.605	0.757	0.600	0.746	1.000

Threshold for multicolinearity $= 0.9$.

- Based on the R output on this slide and the previous one, would you say that the model considered is appropriate? Justify.

No, appropriate since all parameters are not significantly different from zero. There is multicolinearity between models.

Recording of the variables

- Let's center the variables around the mean:

$$(x_i - \bar{x}_i)$$

high correlation is gone

- The correlation matrix of the recoded variables is:

	Y	x_1	x_2	x_1^2	x_2^2	x_1x_2
Y	1.000	-0.556	0.751	0.165	-0.022	0.093
x_1	-0.556	1.000	0.000	0.000	0.000	0.000
x_2	0.751	0.000	1.000	0.000	0.000	0.000
x_1^2	0.165	0.000	0.000	1.000	0.267	0.000
x_2^2	-0.022	0.000	0.000	0.267	1.000	0.000
x_1x_2	0.093	0.000	0.000	0.000	0.000	1.000

Centering good for polynomial models.

R codes

```
attach(Powercell)
x1 = X1 - mean(X1)
x2 = X2 - mean(X2)
fit2 = lm(Y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1*x2))
summary(fit2)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1 * x2))
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
## -21.465   9.263  12.202  41.930  -5.842 -31.842  21.158 -25.404 -20.465   7.263
##      11
##  13.202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  162.8421    16.6076   9.805 0.000188 ***
## x1          -139.5833    33.0418  -4.224 0.008292 **
## x2             7.5500     1.3217   5.712 0.002297 **
## I(x1^2)      171.2171   127.1255   1.347 0.235856
## I(x2^2)       -0.1061     0.2034  -0.521 0.624352
## I(x1 * x2)     2.8750     4.0468   0.710 0.509184
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.37 on 5 degrees of freedom
## Multiple R-squared:  0.9135, Adjusted R-squared:  0.8271
## F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

polynomial terms still not significant

Polynomial regression model and centered data

- Reason of centering: a term and its higher order are highly correlated, centering reduces computation difficulties that may arise.

Hierarchical approach to fitting

- First fit a second-order or third-order model and then explore whether a lower-order model is adequate
- **Exercise 1**:

Consider the third order model with one value

T-test
$\beta_{111}=0$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \beta_{111} X_i^3 + \epsilon_i$$

F-test
β_{11}, β_{111}

How can we test whether the cubic term can be dropped? And how can we test whether both the cubic term and quadratic term can be dropped?

Regression function in terms of the initial variables

- We often wish to express the final model in terms of the original variables (rather than the centered variables).
- Example: we consider the fitted model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_{11} x_i^2 + \hat{\beta}_{111} x_i^3$$

  which we want to express in terms of Xi rather than $x_i = X_i - \bar{X}$.
- **Exercise 2**: Show that the fitted model $x_i = X_i + \bar{X}$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_{11} x_i^2$$

  can be express in terms of X_i as

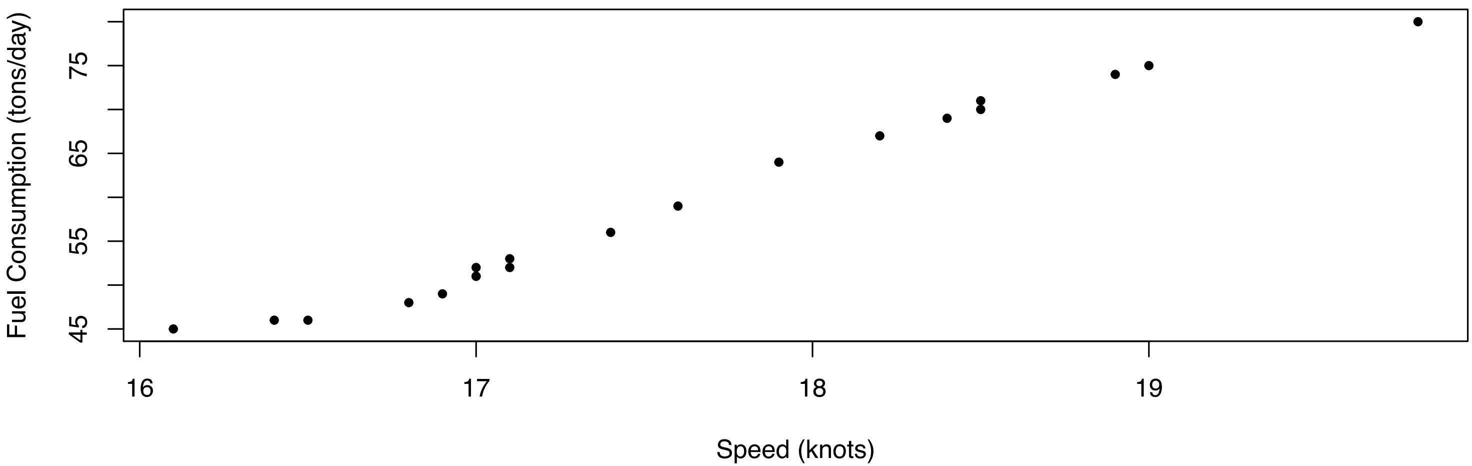
$$\hat{Y}_i = \hat{\beta}'_0 + \hat{\beta}'_1 X_i + \hat{\beta}'_{11} X_i^2$$

Example: Relationship Between Container Ship Speed and Fuel Consumption

Wang and Meng (2012) studied the relationship between Container Ship speed (X, in knots) and fuel consumption (Y , in tons/day)

```
spdfuel=read.csv("ship_speed_fuel.csv", header=T)
attach(spdfuel)

plot(speed,fuel,xlab="Speed (knots)",ylab="Fuel Consumption (tons/day)", pch=20)
```



R codes

$Y = B_0 + B_1 X_1^* + B_{11} X_1^{*2} + B_{111} X_1^{*3}$

```
speed.star = speed - mean(speed)

fit1 = lm(fuel~speed.star + I(speed.star^2)+ I(speed.star^3), data=spdfuel)
summary(fit1)
```

```
##
## Call:
## lm(formula = fuel ~ speed.star + I(speed.star^2) + I(speed.star^3),
##     data = spdfuel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09704 -0.43998 -0.09629  0.47461  1.32907
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       58.7020     0.2324 252.566  < 2e-16 ***
## speed.star        13.3245     0.2993  44.518  < 2e-16 ***
## I(speed.star^2)    0.7779     0.2152   3.616  0.00232 **
## I(speed.star^3)   -1.1479     0.1384  -8.294 3.46e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7171 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9959
## F-statistic:  1551 on 3 and 16 DF,  p-value: < 2.2e-16
```

R codes

```
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: fuel
##                  Df  Sum Sq Mean Sq   F value    Pr(>F)
## speed.star        1 2355.43 2355.43 4580.2738 < 2.2e-16 ***
## I(speed.star^2)  1    2.77    2.77    5.3784   0.03394 *
## I(speed.star^3)  1   35.37   35.37   68.7881 3.462e-07 ***
## Residuals        16    8.23    0.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit2 = lm(fuel~speed.star)
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: fuel
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## speed.star  1 2355.43 2355.43  914.36 < 2.2e-16 ***
## Residuals  18   46.37    2.58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```