# STAC67: Regression Analysis

## Lecture 18

Thibault Randrianarisoa

November 13, 2026

University of Toronto Scarborough

# Review of Wednesday

- Do the groups share one regression function? An $F$ test compares one regression for all the groups with a separate regression in each group (the indicators and all their products with the predictors).
- Polynomial regression is multiple regression, with powers and products as extra columns of $\boldsymbol{X}$. Centre the predictors, prefer low orders, do not extrapolate.
- A product $X_1 X_2$: the slope of one predictor changes linearly with the other.

# Example: power cells

A preliminary experiment on a new type of power cell. Response $Y$: the number of discharge-charge cycles before the cell failed. Predictors: the charge rate $X_1$ (amperes) and the temperature $X_2$ (degrees Celsius), each at three levels.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 150 | 86 | 49 | 288 | 157 | 131 | 184 | 109 | 279 | 235 | 224 |
| $X_1$ | 0.6 | 1.0 | 1.4 | 0.6 | 1.0 | 1.0 | 1.0 | 1.4 | 0.6 | 1.0 | 1.4 |
| $X_2$ | 10 | 10 | 10 | 20 | 20 | 20 | 20 | 20 | 30 | 30 | 30 |

# Example: power cells

A preliminary experiment on a new type of power cell. Response $Y$: the number of discharge-charge cycles before the cell failed. Predictors: the charge rate $X_1$ (amperes) and the temperature $X_2$ (degrees Celsius), each at three levels.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y$ | 150 | 86 | 49 | 288 | 157 | 131 | 184 | 109 | 279 | 235 | 224 |
| $X_1$ | 0.6 | 1.0 | 1.4 | 0.6 | 1.0 | 1.0 | 1.0 | 1.4 | 0.6 | 1.0 | 1.4 |
| $X_2$ | 10 | 10 | 10 | 20 | 20 | 20 | 20 | 20 | 30 | 30 | 30 |

Each of the $3 \times 3 = 9$ combinations of a charge rate and a temperature was tested on one cell, except the middle one, $(X_1, X_2) = (1.0, 20)$, which was tested on three cells: $n = 9 + 2 = 11$.

The second-order model has $p = 6$ parameters, so $n - p = 5$ degrees of freedom are left to estimate $\sigma^2$.

# The second-order model on the raw predictors

`lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2))`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 337.7215 | 149.9616 | 2.2521 | 0.0741 |
| X1 | -539.5175 | 268.8603 | -2.0067 | 0.1011 |
| X2 | 8.9171 | 9.1825 | 0.9711 | 0.3761 |
| I(X1^2) | 171.2171 | 127.1255 | 1.3468 | 0.2359 |
| I(X2^2) | -0.1061 | 0.2034 | -0.5214 | 0.6244 |
| I(X1 * X2) | 2.8750 | 4.0468 | 0.7104 | 0.5092 |

$R^2 = 0.914$, $F^* = 10.57$ on 5 and 5 df, $p = 0.011$.

# The second-order model on the raw predictors

`lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2))`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 337.7215 | 149.9616 | 2.2521 | 0.0741 |
| X1 | -539.5175 | 268.8603 | -2.0067 | 0.1011 |
| X2 | 8.9171 | 9.1825 | 0.9711 | 0.3761 |
| I(X1^2) | 171.2171 | 127.1255 | 1.3468 | 0.2359 |
| I(X2^2) | -0.1061 | 0.2034 | -0.5214 | 0.6244 |
| I(X1 * X2) | 2.8750 | 4.0468 | 0.7104 | 0.5092 |

$R^2 = 0.914$, $F^* = 10.57$ on 5 and 5 df, $p = 0.011$.

The model as a whole is significant, but no single coefficient is. A sign of multicollinearity.

# The second-order model on the raw predictors

```
lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(X1 * X2)):
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  337.7215   149.9616  2.2521   0.0741
X1          -539.5175   268.8603 -2.0067   0.1011
X2             8.9171     9.1825  0.9711   0.3761
I(X1^2)      171.2171   127.1255  1.3468   0.2359
I(X2^2)       -0.1061     0.2034 -0.5214   0.6244
I(X1 * X2)     2.8750     4.0468  0.7104   0.5092
```

$R^2 = 0.914$, $F^* = 10.57$ on 5 and 5 df, $p = 0.011$.

The model as a whole is significant, but no single coefficient is. A sign of multicollinearity.

Indeed $r(X_1, X_1^2) = 0.991$ and $r(X_2, X_2^2) = 0.986$: for $X_1$, the factor $1/(1 - r^2)$ of Lecture 16 is already $56$.

# Centring, and the centred fit

Centre at the means: $x_1 = X_1 - 1$, $x_2 = X_2 - 20$. The correlation between the linear term and its square drops from $0.991$ to $0.000$, for both predictors. The same model on `x1` and `x2`:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  162.8421    16.6076  9.8053   0.0002
x1          -139.5833    33.0418 -4.2245   0.0083
x2            7.5500     1.3217  5.7125   0.0023
I(x1^2)     171.2171   127.1255  1.3468   0.2359
I(x2^2)      -0.1061     0.2034 -0.5214   0.6244
I(x1 * x2)    2.8750     4.0468  0.7104   0.5092
```

# Centring, and the centred fit

Centre at the means: $x_1 = X_1 - 1$, $x_2 = X_2 - 20$. The correlation between the linear term and its square drops from $0.991$ to $0.000$, for both predictors. The same model on `x1` and `x2`:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  162.8421    16.6076  9.8053   0.0002
x1          -139.5833    33.0418 -4.2245   0.0083
x2            7.5500     1.3217  5.7125   0.0023
I(x1^2)     171.2171   127.1255  1.3468   0.2359
I(x2^2)      -0.1061     0.2034 -0.5214   0.6244
I(x1 * x2)    2.8750     4.0468  0.7104   0.5092
```

- The linear terms are now clearly significant ($p = 0.008$ and $0.002$): they are the slopes at the centre of the design.
- The second-order coefficients and their tests, the fitted values and $R^2 = 0.914$ are unchanged.

# Centring, and the centred fit

Centre at the means: $x_1 = X_1 - 1$, $x_2 = X_2 - 20$. The correlation between the linear term and its square drops from $0.991$ to $0.000$, for both predictors. The same model on `x1` and `x2`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 162.8421 | 16.6076 | 9.8053 | 0.0002 |
| x1 | -139.5833 | 33.0418 | -4.2245 | 0.0083 |
| x2 | 7.5500 | 1.3217 | 5.7125 | 0.0023 |
| I(x1^2) | 171.2171 | 127.1255 | 1.3468 | 0.2359 |
| I(x2^2) | -0.1061 | 0.2034 | -0.5214 | 0.6244 |
| I(x1 * x2) | 2.8750 | 4.0468 | 0.7104 | 0.5092 |

- The linear terms are now clearly significant ($p = 0.008$ and $0.002$): they are the slopes at the centre of the design.
- The second-order coefficients and their tests, the fitted values and $R^2 = 0.914$ are unchanged.

Are the three second-order terms needed at all? Need a test with full and reduced models ($F^* = 0.78$ on 3 and 5 df).

# Transforming the response

# Reading a model with a log response

Survival time $Y$ (days) of 54 patients after a liver operation, on four scores. As in Lecture 8 and Assignment 2, Box-Cox gives $\hat{\lambda} = 0.03$, with 95% interval $(-0.25, 0.30)$: take $\ln Y$. `lm(log(Y) ~ X1 + X2 + X3 + X4)`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.8519 | 0.2663 | 14.4666 | 0.0000 |
| X1 | 0.0837 | 0.0288 | 2.9042 | 0.0055 |
| X2 | 0.0127 | 0.0023 | 5.4736 | 0.0000 |
| X3 | 0.0156 | 0.0021 | 7.4403 | 0.0000 |
| X4 | 0.0321 | 0.0515 | 0.6229 | 0.5363 |

# Reading a model with a log response

Survival time $Y$ (days) of 54 patients after a liver operation, on four scores. As in Lecture 8 and Assignment 2, Box-Cox gives $\hat{\lambda} = 0.03$, with 95% interval $(-0.25, 0.30)$: take $\ln Y$. `lm(log(Y) ~ X1 + X2 + X3 + X4)`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.8519 | 0.2663 | 14.4666 | 0.0000 |
| X1 | 0.0837 | 0.0288 | 2.9042 | 0.0055 |
| X2 | 0.0127 | 0.0023 | 5.4736 | 0.0000 |
| X3 | 0.0156 | 0.0021 | 7.4403 | 0.0000 |
| X4 | 0.0321 | 0.0515 | 0.6229 | 0.5363 |

One more unit of $X_k$, the others held fixed, adds $b_k$ to $\ln Y$: it multiplies $Y$ by $e^{b_k}$, a change of $100\,(e^{b_k} - 1)\,\%$.

- Enzyme score: $e^{0.0156} = 1.016$, survival 1.6% longer per point; 95% interval from $e^{0.0114}$ to $e^{0.0198}$: 1.1% to 2.0%.
- Blood clotting score: $e^{0.0837} = 1.087$, 8.7% per point.

# Reading a model with a log response

Survival time $Y$ (days) of 54 patients after a liver operation, on four scores. As in Lecture 8 and Assignment 2, Box-Cox gives $\hat{\lambda} = 0.03$, with 95% interval $(-0.25, 0.30)$: take $\ln Y$. `lm(log(Y) ~ X1 + X2 + X3 + X4)`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 3.8519 | 0.2663 | 14.4666 | 0.0000 |
| X1 | 0.0837 | 0.0288 | 2.9042 | 0.0055 |
| X2 | 0.0127 | 0.0023 | 5.4736 | 0.0000 |
| X3 | 0.0156 | 0.0021 | 7.4403 | 0.0000 |
| X4 | 0.0321 | 0.0515 | 0.6229 | 0.5363 |

One more unit of $X_k$, the others held fixed, adds $b_k$ to $\ln Y$: it multiplies $Y$ by $e^{b_k}$, a change of $100\,(e^{b_k} - 1)\,\%$.

- Enzyme score: $e^{0.0156} = 1.016$, survival 1.6% longer per point; 95% interval from $e^{0.0114}$ to $e^{0.0198}$: 1.1% to 2.0%.
- Blood clotting score: $e^{0.0837} = 1.087$, 8.7% per point.

What is multiplied is the median of $Y$ (Assignment 2, Question 5): $e^{\hat{Y}'}$, with $\hat{Y}'$ the fitted value of $\ln Y$, estimates the median survival time, not the mean.

# Transforming a predictor

# Transform $Y$, or transform $X$?

Two different tools:

- Transforming $Y$ changes the shape of the relationship and the distribution of the errors: their spread and their skewness (asymmetry: a longer tail on one side). It is the remedy for unequal variances and non-normal errors.
- Transforming $X$ changes only the shape of the relationship. The errors stay as they are.

# Transform $Y$, or transform $X$?

Two different tools:

- Transforming $Y$ changes the shape of the relationship and the distribution of the errors: their spread and their skewness (asymmetry: a longer tail on one side). It is the remedy for unequal variances and non-normal errors.
- Transforming $X$ changes only the shape of the relationship. The errors stay as they are.

So when the residuals have a constant spread and look normal, but curve when plotted against a predictor: transform that predictor, or add its square.

# Transform $Y$, or transform $X$?

Two different tools:

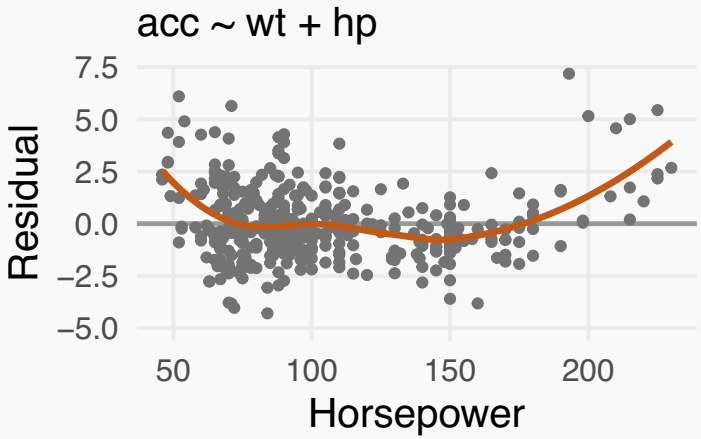
- Transforming $Y$ changes the shape of the relationship and the distribution of the errors: their spread and their skewness (asymmetry: a longer tail on one side). It is the remedy for unequal variances and non-normal errors.
- Transforming $X$ changes only the shape of the relationship. The errors stay as they are.

So when the residuals have a constant spread and look normal, but curve when plotted against a predictor: transform that predictor, or add its square.

Plot the residuals against each predictor, not only against the fitted values. Usual choices: $\ln X$, $\sqrt{X}$, $1/X$.

# Example: acceleration and horsepower

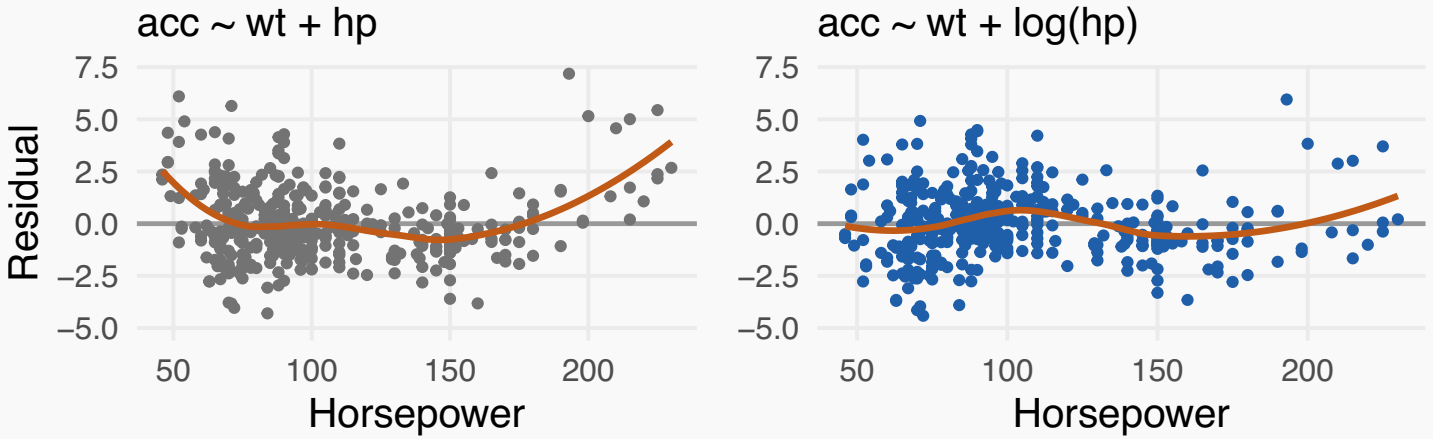
`acc ~ wt + hp`, our model since Lecture 9 (`acc`: seconds from 0 to 60 mph). Residuals against horsepower, with a lowess curve:




The spread is roughly constant, but the residuals are positive at both ends and negative in the middle: the effect of horsepower is not linear.

# Example: acceleration and horsepower

`acc ~ wt + hp`, our model since Lecture 9 (`acc`: seconds from 0 to 60 mph). Residuals against horsepower, with a lowess curve:




The spread is roughly constant, but the residuals are positive at both ends and negative in the middle: the effect of horsepower is not linear.

With `log(hp)` in place of `hp`, most of the curve is gone, and $R^2$ goes from 0.604 to 0.654 with the same number of parameters.

# Reading a predictor in logs

$$\widehat{\mathsf{acc}} = 60.33 + 0.0027\,\mathsf{wt} - 11.53\,\ln(\mathsf{hp}).$$

Let $b$ be the coefficient of $\ln X$. Multiplying $X$ by $c$ adds $b \ln c$ to $\hat{Y}$, whatever the starting value of $X$, the other predictors held fixed.

# Reading a predictor in logs

$$\widehat{\texttt{acc}} = 60.33 + 0.0027\,\texttt{wt} - 11.53\,\ln(\texttt{hp}).$$

Let $b$ be the coefficient of $\ln X$. Multiplying $X$ by $c$ adds $b \ln c$ to $\hat{Y}$, whatever the starting value of $X$, the other predictors held fixed.

- 1% more horsepower: $b \ln 1.01 \approx b/100 = -0.115$ s.
- 10% more: $b \ln 1.1 = -1.10$ s. Twice the horsepower: $b \ln 2 = -7.99$ s.

# Reading a predictor in logs

$$\widehat{\mathsf{acc}} = 60.33 + 0.0027\,\mathsf{wt} - 11.53\,\ln(\mathsf{hp}).$$

Let $b$ be the coefficient of $\ln X$. Multiplying $X$ by $c$ adds $b \ln c$ to $\hat{Y}$, whatever the starting value of $X$, the other predictors held fixed.

- 1% more horsepower: $b \ln 1.01 \approx b/100 = -0.115$ s.
- 10% more: $b \ln 1.1 = -1.10$ s. Twice the horsepower: $b \ln 2 = -7.99$ s.

The effect is the same for equal *ratios*, not for equal differences. Ten more horsepower are worth $-1.78$ s from 60 to 70 hp, and only $-0.56$ s from 200 to 210 hp: diminishing returns.

# Logs on either side: how to read $b$

| Response | Predictor | Reading of its coefficient $b$ |
|---|---|---|
| $Y$ | $X$ | one more unit of $X$: $Y$ changes by $b$ units (on average) |
| $\ln Y$ | $X$ | one more unit of $X$: $Y$ times $e^b$, about $100\,b\,\%$ more (on the median) |
| $Y$ | $\ln X$ | 1% more $X$: $Y$ changes by about $b/100$ units (on average) |
| $\ln Y$ | $\ln X$ | 1% more $X$: $Y$ changes by about $b\,\%$ (on the median) |

# Logs on either side: how to read $b$

| Response | Predictor | Reading of its coefficient $b$ |
|---|---|---|
| $Y$ | $X$ | one more unit of $X$: $Y$ changes by $b$ units |
| $\ln Y$ | $X$ | one more unit of $X$: $Y$ times $e^b$, about $100\,b\,\%$ more |
| $Y$ | $\ln X$ | 1% more $X$: $Y$ changes by about $b/100$ units |
| $\ln Y$ | $\ln X$ | 1% more $X$: $Y$ changes by about $b\,\%$ |

- Always with the other predictors held fixed; with $\ln Y$, the statement is about the median of $Y$.
- The approximations are for small changes. Exactly: $e^b - 1$ in the second row, and $X$ times $c$ multiplies $Y$ by $c^b$ in the last one.

# Logs on either side: how to read $b$

| Response | Predictor | Reading of its coefficient $b$ |
|---|---|---|
| $Y$ | $X$ | one more unit of $X$: $Y$ changes by $b$ units |
| $\ln Y$ | $X$ | one more unit of $X$: $Y$ times $e^b$, about $100\,b\,\%$ more |
| $Y$ | $\ln X$ | 1% more $X$: $Y$ changes by about $b/100$ units |
| $\ln Y$ | $\ln X$ | 1% more $X$: $Y$ changes by about $b\,\%$ |

- Always with the other predictors held fixed; with $\ln Y$, the statement is about the median of $Y$.
- The approximations are for small changes. Exactly: $e^b - 1$ in the second row, and $X$ times $c$ multiplies $Y$ by $c^b$ in the last one.

The last row is a power law, $Y \propto X^b$: for instance the fuel burnt by a ship against its speed (problem set).

# A log, or a square?

Wednesday's remedy for a curve was a square. Three models for the acceleration:

| Model | $p$ | $R^2$ | adjusted $R^2$ |
|---|---|---|---|
| `acc ~ wt + hp` | 3 | 0.604 | 0.602 |
| `acc ~ wt + log(hp)` | 3 | 0.654 | 0.652 |
| `acc ~ wt + hp + I(hp^2)` | 4 | 0.665 | 0.662 |

# A log, or a square?

Wednesday's remedy for a curve was a square. Three models for the acceleration:

| Model | $p$ | $R^2$ | adjusted $R^2$ |
|---|---|---|---|
| `acc ~ wt + hp` | 3 | 0.604 | 0.602 |
| `acc ~ wt + log(hp)` | 3 | 0.654 | 0.652 |
| `acc ~ wt + hp + I(hp^2)` | 4 | 0.665 | 0.662 |

- The square is more flexible and fits slightly better. But there are two coefficients to read, and the fitted parabola turns at 256 hp, just beyond the data (the maximum is 230).
- The log has one coefficient, read in per cent, and never turns.

# A log, or a square?

Wednesday's remedy for a curve was a square. Three models for the acceleration:

| Model | $p$ | $R^2$ | adjusted $R^2$ |
|---|---|---|---|
| `acc ~ wt + hp` | 3 | 0.604 | 0.602 |
| `acc ~ wt + log(hp)` | 3 | 0.654 | 0.652 |
| `acc ~ wt + hp + I(hp^2)` | 4 | 0.665 | 0.662 |

- The square is more flexible and fits slightly better. But there are two coefficients to read, and the fitted parabola turns at 256 hp, just beyond the data (the maximum is 230).
- The log has one coefficient, read in per cent, and never turns.

The last two models are not nested: no $F$ test can compare them. We need criteria to compare any two models. This is the rest of today.

# Model selection

# Which predictors?

With $P - 1$ candidate predictors (squares, products and indicators included), each one in or out, there are $2^{P-1}$ possible models: 16 for 4 predictors, 256 for 8, more than a million for 20.

# Which predictors?

With $P - 1$ candidate predictors (squares, products and indicators included), each one in or out, there are $2^{P-1}$ possible models: 16 for 4 predictors, 256 for 8, more than a million for 20.

- Too few predictors: one that matters is left out, and the coefficients of the others absorb its effect, as for the iris flowers without their species (Simpson's paradox, Lecture 16). The fitted values are biased.
- Too many: the variances inflate (multicollinearity), the model follows the noise of this sample, and it is harder to read and to use.

# Which predictors?

With $P - 1$ candidate predictors (squares, products and indicators included), each one in or out, there are $2^{P-1}$ possible models: 16 for 4 predictors, 256 for 8, more than a million for 20.

- Too few predictors: one that matters is left out, and the coefficients of the others absorb its effect, as for the iris flowers without their species (Simpson's paradox, Lecture 16). The fitted values are biased.
- Too many: the variances inflate (multicollinearity), the model follows the noise of this sample, and it is harder to read and to use.

The right balance depends on the aim: to explain (each coefficient is read, so the model must stay readable) or to predict (only the quality of $\hat{Y}$ for new cases matters).

# Two steps, and the data

1. Model selection: narrow the possible models down to one or a few good ones. Today.
2. Model validation: check that the chosen model still does well on data that were not used to choose it. Wednesday.

# Two steps, and the data

1. Model selection: narrow the possible models down to one or a few good ones. Today.
2. Model validation: check that the chosen model still does well on data that were not used to choose it. Wednesday.

The surgical unit: 108 patients, split at random into 54 to build the model (the data of the log response) and 54 kept aside for validation. Response: $\ln Y$. Candidate predictors:

- the four scores: blood clotting $X_1$, prognostic index $X_2$, enzyme $X_3$, liver function $X_4$;
- age $X_5$, gender $X_6$, and alcohol use, a factor with three levels (none, moderate, severe), through the indicators $X_7$ and $X_8$.

# Two steps, and the data

1. Model selection: narrow the possible models down to one or a few good ones. Today.
2. Model validation: check that the chosen model still does well on data that were not used to choose it. Wednesday.

The surgical unit: 108 patients, split at random into 54 to build the model (the data of the log response) and 54 kept aside for validation. Response: $\ln Y$. Candidate predictors:

- the four scores: blood clotting $X_1$, prognostic index $X_2$, enzyme $X_3$, liver function $X_4$;
- age $X_5$, gender $X_6$, and alcohol use, a factor with three levels (none, moderate, severe), through the indicators $X_7$ and $X_8$.

We start with $X_1$ to $X_4$ only: 16 models.

# Why $R^2$ cannot choose

A candidate model has $p$ parameters ($p-1$ predictors and the intercept) and an error sum of squares $SSE_p$. $SST$ is the same for all the models.

$$R^2_p = 1 - \frac{SSE_p}{SST}.$$

# Why $R^2$ cannot choose

A candidate model has $p$ parameters ($p-1$ predictors and the intercept) and an error sum of squares $SSE_p$. $SST$ is the same for all the models.

$$R^2_p = 1 - \frac{SSE_p}{SST}.$$

Adding a predictor never increases $SSE_p$ (Lecture 12), so $R^2_p$ never decreases: the largest $R^2_p$ always belongs to the model with all the predictors.

# Why $R^2$ cannot choose

A candidate model has $p$ parameters ($p-1$ predictors and the intercept) and an error sum of squares $SSE_p$. $SST$ is the same for all the models.

$$R_p^2 = 1 - \frac{SSE_p}{SST}.$$

Adding a predictor never increases $SSE_p$ (Lecture 12), so $R_p^2$ never decreases: the largest $R_p^2$ always belongs to the model with all the predictors.

$R_p^2$ can compare models of the same size. Across sizes, plot the best $R_p^2$ of each size against $p$: where the curve flattens, more predictors add little. The other criteria make this precise: each extra parameter has a price, and a larger model wins only if it improves the fit by more than that price.

# Adjusted $R^2$

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SST} = 1 - \frac{MSE_p}{SST/(n-1)}.$$

$SST/(n-1)$ is the same for every model: the largest $R^2_{a,p}$ is the smallest $MSE_p$.

# Adjusted $R^2$

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SST} = 1 - \frac{MSE_p}{SST/(n-1)}.$$

$SST/(n-1)$ is the same for every model: the largest $R^2_{a,p}$ is the smallest $MSE_p$.

Adding a predictor lowers $SSE_p$, but also $n-p$: $MSE_p = SSE_p/(n-p)$ can go up, and $R^2_{a,p}$ down.

# Adjusted $R^2$

$$R^2_{a,p} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SST} = 1 - \frac{MSE_p}{SST/(n-1)}.$$

$SST/(n-1)$ is the same for every model: the largest $R^2_{a,p}$ is the smallest $MSE_p$.

Adding a predictor lowers $SSE_p$, but also $n-p$: $MSE_p = SSE_p/(n-p)$ can go up, and $R^2_{a,p}$ down.

How high is the price? Adding one predictor raises $R^2_a$ exactly when its $t$ statistic has $|t^*| > 1$ (next week's problem set). This is a low bar: adjusted $R^2$ tends to keep too many predictors.

# Mallows' $C_p$: the idea

Let $\mu_i = E(Y_i)$ be the true mean of case $i$. The fitted value $\hat{Y}_i$ of a candidate model can miss it in two ways:

$$E\big(\hat{Y}_i - \mu_i\big)^2 = \underbrace{\big(E(\hat{Y}_i) - \mu_i\big)^2}_{\text{bias}_i^2\text{: predictors left out}} + \underbrace{Var(\hat{Y}_i)}_{\text{parameters estimated}}$$

# Mallows' $C_p$: the idea

Let $\mu_i = E(Y_i)$ be the true mean of case $i$. The fitted value $\hat{Y}_i$ of a candidate model can miss it in two ways:

$$E\big(\hat{Y}_i - \mu_i\big)^2 = \underbrace{\big(E(\hat{Y}_i) - \mu_i\big)^2}_{\text{bias}_i^2\text{: predictors left out}} + \underbrace{Var(\hat{Y}_i)}_{\text{parameters estimated}}$$

Sum over the $n$ cases, in units of $\sigma^2$: the total mean squared error of the model,

$$\Gamma_p = \frac{1}{\sigma^2}\Big[\sum_i \text{bias}_i^2 + \sum_i Var(\hat{Y}_i)\Big] = \frac{\sum_i \text{bias}_i^2}{\sigma^2} + p,$$

since $\sum_i Var(\hat{Y}_i) = \sigma^2 \sum_i h_{ii} = \sigma^2\, tr(\boldsymbol{H}_p) = p\,\sigma^2$ (Lecture 12).

# Mallows' $C_p$: the idea

Let $\mu_i = E(Y_i)$ be the true mean of case $i$. The fitted value $\hat{Y}_i$ of a candidate model can miss it in two ways:

$$E\big(\hat{Y}_i - \mu_i\big)^2 = \underbrace{\big(E(\hat{Y}_i) - \mu_i\big)^2}_{\text{bias}_i^2\text{: predictors left out}} + \underbrace{Var(\hat{Y}_i)}_{\text{parameters estimated}}$$

Sum over the $n$ cases, in units of $\sigma^2$: the total mean squared error of the model,

$$\Gamma_p = \frac{1}{\sigma^2}\Big[\sum_i \text{bias}_i^2 + \sum_i Var(\hat{Y}_i)\Big] = \frac{\sum_i \text{bias}_i^2}{\sigma^2} + p,$$

since $\sum_i Var(\hat{Y}_i) = \sigma^2 \sum_i h_{ii} = \sigma^2\, tr(\boldsymbol{H}_p) = p\,\sigma^2$ (Lecture 12).

The trade-off in one line: a new predictor can only reduce the bias, and it always costs 1 (in the variance term).

# Mallows' $C_p$: the estimate

By the mean of a quadratic form (Lecture 12), whether or not the model is right:

$$E(SSE_p) = E\big(\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H}_p)\boldsymbol{Y}\big) = (n-p)\,\sigma^2 + \sum_i \text{bias}_i^2.$$

# Mallows' $C_p$: the estimate

By the mean of a quadratic form (Lecture 12), whether or not the model is right:

$$E(SSE_p) = E\big(\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H}_p)\boldsymbol{Y}\big) = (n-p)\,\sigma^2 + \sum_i \text{bias}_i^2.$$

Hence $\Gamma_p = E(SSE_p)/\sigma^2 - (n-2p)$. Replace $E(SSE_p)$ by $SSE_p$, and $\sigma^2$ by the $MSE$ of the model with all $P-1$ predictors, taken to have no bias:

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n-2p).$$

# Mallows' $C_p$: the estimate

By the mean of a quadratic form (Lecture 12), whether or not the model is right:

$$E(SSE_p) = E\big(\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H}_p)\boldsymbol{Y}\big) = (n-p)\,\sigma^2 + \sum_i \text{bias}_i^2.$$

Hence $\Gamma_p = E(SSE_p)/\sigma^2 - (n-2p)$. Replace $E(SSE_p)$ by $SSE_p$, and $\sigma^2$ by the $MSE$ of the model with all $P-1$ predictors, taken to have no bias:

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n-2p).$$

- Look for a small $C_p$: a small total error.
- A model with no bias has $C_p$ close to $p$. A $C_p$ far above $p$ signals bias.
- For the model with all the predictors, $C_P = P$ exactly: it says nothing there.

# AIC and BIC

Akaike's information criterion and the Bayesian information criterion (also called SBC):

$$AIC_p = n \ln\left(\frac{SSE_p}{n}\right) + 2p, \qquad BIC_p = n \ln\left(\frac{SSE_p}{n}\right) + p \ln n.$$

Small is good. The first term falls when a predictor is added (better fit); the second rises (the price).

# AIC and BIC

Akaike's information criterion and the Bayesian information criterion (also called SBC):

$$AIC_p = n \ln\left(\frac{SSE_p}{n}\right) + 2p, \qquad BIC_p = n \ln\left(\frac{SSE_p}{n}\right) + p \ln n.$$

Small is good. The first term falls when a predictor is added (better fit); the second rises (the price).

- $\ln n > 2$ as soon as $n \geq 8$: BIC charges more per parameter, and chooses smaller models than AIC.
- Both come from the likelihood of the normal error model: minus twice its maximum, plus a penalty. Their values mean nothing alone; only differences between models count.

# AIC and BIC

Akaike's information criterion and the Bayesian information criterion (also called SBC):

$$AIC_p = n \ln\left(\frac{SSE_p}{n}\right) + 2p, \qquad BIC_p = n \ln\left(\frac{SSE_p}{n}\right) + p \ln n.$$

Small is good. The first term falls when a predictor is added (better fit); the second rises (the price).

- $\ln n > 2$ as soon as $n \geq 8$: BIC charges more per parameter, and chooses smaller models than AIC.
- Both come from the likelihood of the normal error model: minus twice its maximum, plus a penalty. Their values mean nothing alone; only differences between models count.

In R, `extractAIC(fit)` gives $AIC_p$ as above. `AIC(fit)` and `BIC(fit)` add a constant that is the same for every model: same ranking.

# PRESS: predicting each case without it

Delete case $i$, fit the model on the other $n-1$ cases, and predict $Y_i$: call it $\hat{Y}_{i(i)}$. The prediction sum of squares is

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)}\right)^2.$$

Small is good. Each case is predicted by a model that has not seen it when fitted: it measures prediction, where $SSE_p$ measures fit.

# PRESS: predicting each case without it

Delete case $i$, fit the model on the other $n-1$ cases, and predict $Y_i$: call it $\hat{Y}_{i(i)}$. The prediction sum of squares is

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)}\right)^2.$$

Small is good. Each case is predicted by a model that has not seen it when fitted: it measures prediction, where $SSE_p$ measures fit.

No need for $n$ fits. With $h_{ii}$ the leverage of case $i$ (Lectures 11 and 12),

$$Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}, \quad \text{so} \quad PRESS_p = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \geq SSE_p.$$

$e_i = Y_i - \hat{Y}_i$ fitted on $n$ obs.

$\hat{Y}_{i(i)}$: fitted on $n-1$ obs.

$1-h_{ii}$: diagonal of hat matrix $H \in \mathbb{R}^{n \times n}$

$SSE_p = \sum e_i^2$

# PRESS: predicting each case without it

Delete case $i$, fit the model on the other $n-1$ cases, and predict $Y_i$: call it $\hat{Y}_{i(i)}$. The prediction sum of squares is

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)}\right)^2.$$

Small is good. Each case is predicted by a model that has not seen it when fitted: it measures prediction, where $SSE_p$ measures fit.

No need for $n$ fits. With $h_{ii}$ the leverage of case $i$ (Lectures 11 and 12),

$$Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}, \qquad \text{so} \qquad PRESS_p = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2 \geq SSE_p.$$

A case with a high leverage pulls the fit towards itself: its residual $e_i$ favors the model, and the division by $1-h_{ii}$ corrects for it.

# The criteria at a glance

| Criterion | A good model has | Price of one more parameter |
|---|---|---|
| $R_p^2$ | (a large value) | none: cannot compare sizes |
| $R_{a,p}^2$ | the largest value | one error degree of freedom |
| $C_p$ | a small value, close to $p$ | 2, in units of $\sigma^2$ |
| $AIC_p$ | the smallest value | 2 |
| $BIC_p$ | the smallest value | $\ln n$ |
| $PRESS_p$ | the smallest value | implicit: fitting noise predicts worse |

# The criteria at a glance

| Criterion | A good model has | Price of one more parameter |
|---|---|---|
| $R_p^2$ | (a large value) | none: cannot compare sizes |
| $R_{a,p}^2$ | the largest value | one error degree of freedom |
| $C_p$ | a small value, close to $p$ | 2, in units of $\sigma^2$ |
| $AIC_p$ | the smallest value | 2 |
| $BIC_p$ | the smallest value | $\ln n$ |
| $PRESS_p$ | the smallest value | implicit: fitting noise predicts worse |

All of them compare models for the same response on the same cases: not a model for $Y$ with a model for $\ln Y$, and not models fitted on different rows.

# Exercise

Surgical unit, $n = 54$, response $\ln Y$, for which $SST = 12.805$. Two models:

| Predictors | $SSE$ | $\ln(SSE/n)$ |
|---|---|---|
| $X_1, X_2, X_3, X_4$ | 3.084 | $-2.863$ |
| $X_1, X_3$ | 5.783 | $-2.234$ |

1. Compute $R^2$, adjusted $R^2$, $C_p$ and $AIC$ for each model. For $C_p$, the model with all the predictors is the model with the four scores.
2. Which model does each criterion prefer?

# Exercise: solution

The $MSE$ of the model with all the predictors is $3.084/49 = 0.0629$.

| | $X_1, X_2, X_3, X_4$ $(p = 5)$ | $X_1, X_3$ $(p = 3)$ |
|---|---|---|
| $R^2_p$ | $1 - \frac{3.084}{12.805} = 0.759$ | 0.548 |
| $R^2_{a,p}$ | $1 - \frac{53}{49} \times \frac{3.084}{12.805} = 0.739$ | $1 - \frac{53}{51} \times \frac{5.783}{12.805} = 0.531$ |
| $C_p$ | $\frac{3.084}{0.0629} - (54 - 10) = 5.00$ | $\frac{5.783}{0.0629} - (54 - 6) = 43.87$ |
| $AIC_p$ | $54 \times (-2.863) + 10 = -144.59$ | $54 \times (-2.234) + 6 = -114.64$ |

# Exercise: solution

The $MSE$ of the model with all the predictors is $3.084/49 = 0.0629$.

| | $X_1, X_2, X_3, X_4$ $(p = 5)$ | $X_1, X_3$ $(p = 3)$ |
|---|---|---|
| $R^2_p$ | $1 - \frac{3.084}{12.805} = 0.759$ | 0.548 |
| $R^2_{a,p}$ | $1 - \frac{53}{49} \times \frac{3.084}{12.805} = 0.739$ | $1 - \frac{53}{51} \times \frac{5.783}{12.805} = 0.531$ |
| $C_p$ | $\frac{3.084}{0.0629} - (54 - 10) = 5.00$ | $\frac{5.783}{0.0629} - (54 - 6) = 43.87$ |
| $AIC_p$ | $54 \times (-2.863) + 10 = -144.59$ | $54 \times (-2.234) + 6 = -114.64$ |

Every criterion prefers the four scores. The $C_p$ of 43.9 for $p = 3$ says that leaving out $X_2$ and $X_4$ biases the fit badly.

# Exercise: solution

The $MSE$ of the model with all the predictors is $3.084/49 = 0.0629$.

| | $X_1, X_2, X_3, X_4$ $(p = 5)$ | $X_1, X_3$ $(p = 3)$ |
|---|---|---|
| $R^2_p$ | $1 - \frac{3.084}{12.805} = 0.759$ | 0.548 |
| $R^2_{a,p}$ | $1 - \frac{53}{49} \times \frac{3.084}{12.805} = 0.739$ | $1 - \frac{53}{51} \times \frac{5.783}{12.805} = 0.531$ |
| $C_p$ | $\frac{3.084}{0.0629} - (54 - 10) = 5.00$ | $\frac{5.783}{0.0629} - (54 - 6) = 43.87$ |
| $AIC_p$ | $54 \times (-2.863) + 10 = -144.59$ | $54 \times (-2.234) + 6 = -114.64$ |

Every criterion prefers the four scores. The $C_p$ of 43.9 for $p = 3$ says that leaving out $X_2$ and $X_4$ biases the fit badly.

But is there a model between the two that does even better?

# All 16 models: one or two predictors

The criteria for every subset of $X_1, X_2, X_3, X_4$ (continued on the next slide):

| Predictors | $p$ | $R_p^2$ | $R_{a,p}^2$ | $C_p$ | $AIC_p$ | $BIC_p$ | $PRESS_p$ |
|---|---|---|---|---|---|---|---|
| none | 1 | 0.000 | 0.000 | 151.44 | −75.72 | −73.73 | 13.292 |
| $X_1$ | 2 | 0.061 | 0.043 | 141.09 | −77.10 | −73.12 | 13.508 |
| $X_2$ | 2 | 0.221 | 0.206 | 108.47 | −87.21 | −83.23 | 10.738 |
| $X_3$ | 2 | 0.427 | 0.416 | 66.52 | −103.81 | −99.83 | 8.329 |
| $X_4$ | 2 | 0.421 | 0.410 | 67.70 | −103.27 | −99.29 | 8.024 |
| $X_1\ X_2$ | 3 | 0.263 | 0.234 | 101.94 | −88.19 | −82.23 | 11.055 |
| $X_1\ X_3$ | 3 | 0.548 | 0.531 | 43.87 | −114.64 | −108.68 | 6.989 |
| $X_1\ X_4$ | 3 | 0.430 | 0.408 | 67.96 | −102.07 | −96.10 | 8.472 |
| $X_2\ X_3$ | 3 | 0.663 | 0.650 | 20.52 | −130.48 | −124.51 | 5.066 |
| $X_2\ X_4$ | 3 | 0.483 | 0.463 | 57.17 | −107.34 | −101.38 | 7.474 |
| $X_3\ X_4$ | 3 | 0.599 | 0.583 | 33.54 | −121.09 | −115.12 | 6.123 |

# All 16 models: three or four predictors

| Predictors | $p$ | $R^2_p$ | $R^2_{a,p}$ | $C_p$ | $AIC_p$ | $BIC_p$ | $PRESS_p$ |
|---|---|---|---|---|---|---|---|
| $X_1\ X_2\ X_3$ | 4 | 0.757 | 0.743 | 3.39 | −146.16 | −138.21 | 3.914 |
| $X_1\ X_2\ X_4$ | 4 | 0.487 | 0.456 | 58.36 | −105.76 | −97.81 | 7.902 |
| $X_1\ X_3\ X_4$ | 4 | 0.612 | 0.589 | 32.96 | −120.82 | −112.87 | 6.209 |
| $X_2\ X_3\ X_4$ | 4 | 0.718 | 0.701 | 11.43 | −138.01 | −130.05 | 4.598 |
| $X_1\ X_2\ X_3\ X_4$ | 5 | 0.759 | 0.739 | 5.00 | −144.59 | −134.64 | 4.069 |

# All 16 models: three or four predictors

| Predictors | $p$ | $R^2_p$ | $R^2_{a,p}$ | $C_p$ | $AIC_p$ | $BIC_p$ | $PRESS_p$ |
|---|---|---|---|---|---|---|---|
| $X_1\ X_2\ X_3$ | 4 | 0.757 | 0.743 | 3.39 | −146.16 | −138.21 | 3.914 |
| $X_1\ X_2\ X_4$ | 4 | 0.487 | 0.456 | 58.36 | −105.76 | −97.81 | 7.902 |
| $X_1\ X_3\ X_4$ | 4 | 0.612 | 0.589 | 32.96 | −120.82 | −112.87 | 6.209 |
| $X_2\ X_3\ X_4$ | 4 | 0.718 | 0.701 | 11.43 | −138.01 | −130.05 | 4.598 |
| $X_1\ X_2\ X_3\ X_4$ | 5 | 0.759 | 0.739 | 5.00 | −144.59 | −134.64 | 4.069 |

- Every criterion chooses $X_1, X_2, X_3$: the best value of each one, in colour, is on its row. Its $C_p = 3.39$ is below $p = 4$: no sign of bias.
- Adding $X_4$ raises $R^2_p$ a little, as it must, and makes every other criterion worse.

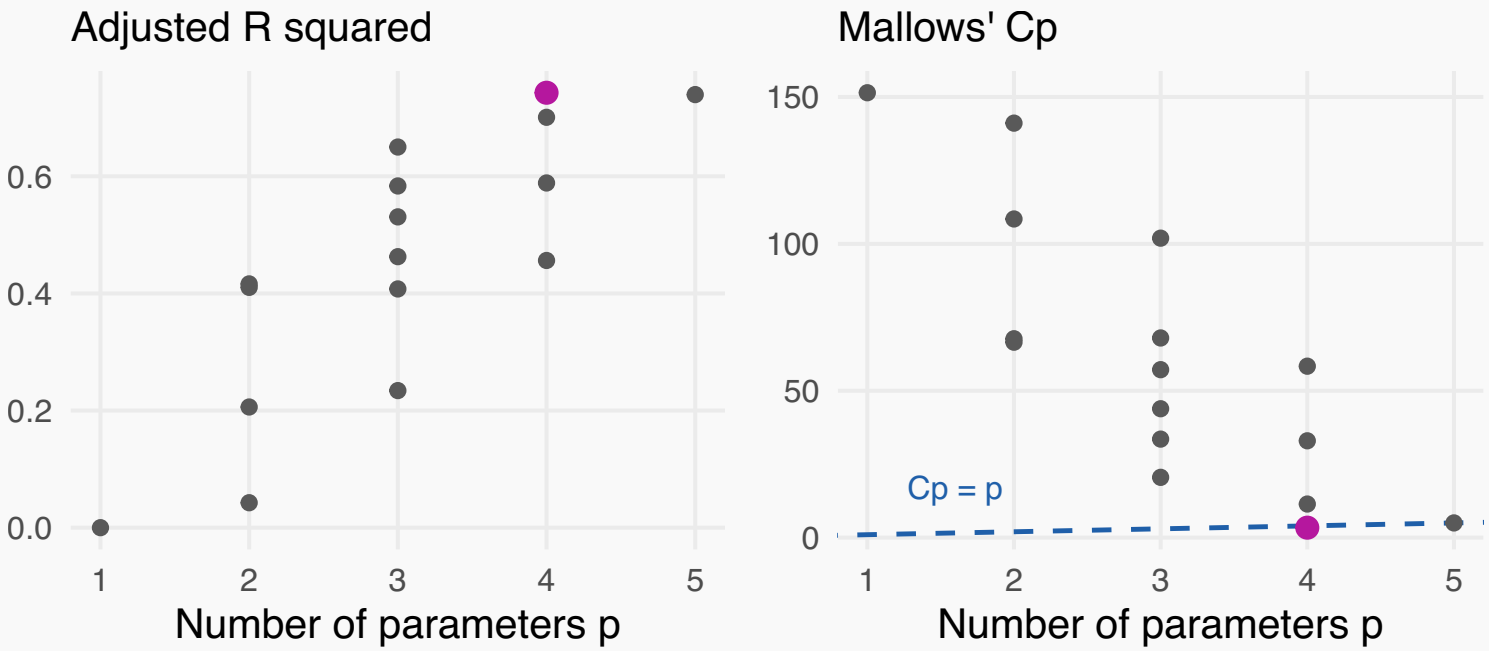
# All 16 models: three or four predictors

| Predictors | $p$ | $R^2_p$ | $R^2_{a,p}$ | $C_p$ | $AIC_p$ | $BIC_p$ | $PRESS_p$ |
|---|---|---|---|---|---|---|---|
| $X_1\ X_2\ X_3$ | 4 | 0.757 | 0.743 | 3.39 | −146.16 | −138.21 | 3.914 |
| $X_1\ X_2\ X_4$ | 4 | 0.487 | 0.456 | 58.36 | −105.76 | −97.81 | 7.902 |
| $X_1\ X_3\ X_4$ | 4 | 0.612 | 0.589 | 32.96 | −120.82 | −112.87 | 6.209 |
| $X_2\ X_3\ X_4$ | 4 | 0.718 | 0.701 | 11.43 | −138.01 | −130.05 | 4.598 |
| $X_1\ X_2\ X_3\ X_4$ | 5 | 0.759 | 0.739 | 5.00 | −144.59 | −134.64 | 4.069 |

- Every criterion chooses $X_1, X_2, X_3$: the best value of each one, in colour, is on its row. Its $C_p = 3.39$ is below $p = 4$: no sign of bias.
- Adding $X_4$ raises $R^2_p$ a little, as it must, and makes every other criterion worse.

Alone, the liver score $X_4$ is the second best predictor ($R^2 = 0.42$). It is correlated with the three others ($r = 0.50$, $0.37$ and $0.42$), and brings nothing once they are in.

# Seeing the 16 models




Each point is a model; $X_1, X_2, X_3$ is in colour. Within a size, every criterion ranks the models as $SSE_p$ does, except PRESS. The criteria matter when sizes are compared.

# In R

For one candidate model, and the model with all the predictors:

```
fit  <- lm(log(Y) ~ X1 + X2 + X3, data = surg)
full <- lm(log(Y) ~ X1 + X2 + X3 + X4, data = surg)
n <- nrow(surg); p <- length(coef(fit))
summary(fit)$adj.r.squared                    # adjusted R2
sum(resid(fit)^2) / sigma(full)^2 - (n - 2 * p)    # Cp
extractAIC(fit)                               # p and AIC
extractAIC(fit, k = log(n))                   # p and BIC
sum((resid(fit) / (1 - hatvalues(fit)))^2)    # PRESS
```

# In R

For one candidate model, and the model with all the predictors:

```
fit  <- lm(log(Y) ~ X1 + X2 + X3, data = surg)
full <- lm(log(Y) ~ X1 + X2 + X3 + X4, data = surg)
n <- nrow(surg); p <- length(coef(fit))
summary(fit)$adj.r.squared                    # adjusted R2
sum(resid(fit)^2) / sigma(full)^2 - (n - 2 * p)    # Cp
extractAIC(fit)                               # p and AIC
extractAIC(fit, k = log(n))                   # p and BIC
sum((resid(fit) / (1 - hatvalues(fit)))^2)    # PRESS
```

This gives 0.743, 3.39, $-146.16$, $-138.21$ and 3.914: the row of the table.

# In R

For one candidate model, and the model with all the predictors:

```
fit  <- lm(log(Y) ~ X1 + X2 + X3, data = surg)
full <- lm(log(Y) ~ X1 + X2 + X3 + X4, data = surg)
n <- nrow(surg); p <- length(coef(fit))
summary(fit)$adj.r.squared                    # adjusted R2
sum(resid(fit)^2) / sigma(full)^2 - (n - 2 * p)    # Cp
extractAIC(fit)                               # p and AIC
extractAIC(fit, k = log(n))                   # p and BIC
sum((resid(fit) / (1 - hatvalues(fit)))^2)    # PRESS
```

This gives 0.743, 3.39, $-146.16$, $-138.21$ and 3.914: the row of the table.

`sigma(full)` is $\sqrt{MSE}$ of the full model; `hatvalues(fit)` gives the leverages $h_{ii}$.

# With the eight predictors: 256 models

The best model of each size (smallest $SSE_p$), with the best value of each criterion over the 256 models in colour:

| Predictors $X_k$ | $p$ | $R^2_{a,p}$ | $C_p$ | $AIC_p$ | $BIC_p$ | $PRESS_p$ |
|---|---|---|---|---|---|---|
| 3 | 2 | 0.416 | 117.48 | −103.81 | −99.83 | 8.329 |
| 2, 3 | 3 | 0.650 | 50.49 | −130.48 | −124.51 | 5.066 |
| 2, 3, 8 | 4 | 0.765 | 18.90 | −151.00 | −143.05 | 3.468 |
| 1, 2, 3, 8 | 5 | 0.816 | 5.73 | −163.38 | −153.43 | 2.736 |
| 1, 2, 3, 6, 8 | 6 | 0.821 | 5.53 | −163.83 | −151.89 | 2.782 |
| 1, 2, 3, 5, 6, 8 | 7 | 0.823 | 5.77 | −163.86 | −149.94 | 2.771 |
| 1, 2, 3, 5, 6, 7, 8 | 8 | 0.823 | 7.03 | −162.74 | −146.83 | 2.808 |
| 1, 2, 3, 4, 5, 6, 7, 8 | 9 | 0.819 | 9.00 | −160.78 | −142.88 | 2.931 |

# With the eight predictors: 256 models

The best model of each size (smallest $SSE_p$), with the best value of each criterion over the 256 models in colour:

| Predictors $X_k$ | $p$ | $R^2_{a,p}$ | $C_p$ | $AIC_p$ | $BIC_p$ | $PRESS_p$ |
|---|---|---|---|---|---|---|
| 3 | 2 | 0.416 | 117.48 | −103.81 | −99.83 | 8.329 |
| 2, 3 | 3 | 0.650 | 50.49 | −130.48 | −124.51 | 5.066 |
| 2, 3, 8 | 4 | 0.765 | 18.90 | −151.00 | −143.05 | 3.468 |
| 1, 2, 3, 8 | 5 | 0.816 | 5.73 | −163.38 | −153.43 | 2.736 |
| 1, 2, 3, 6, 8 | 6 | 0.821 | 5.53 | −163.83 | −151.89 | 2.782 |
| 1, 2, 3, 5, 6, 8 | 7 | 0.823 | 5.77 | −163.86 | −149.94 | 2.771 |
| 1, 2, 3, 5, 6, 7, 8 | 8 | 0.823 | 7.03 | −162.74 | −146.83 | 2.808 |
| 1, 2, 3, 4, 5, 6, 7, 8 | 9 | 0.819 | 9.00 | −160.78 | −142.88 | 2.931 |

The criteria no longer agree: $R^2_a$ and AIC keep 6 predictors, $C_p$ keeps 5, BIC and PRESS keep 4. And the differences between these models are small.

# In R: the best model of each size

The package `leaps` searches all the subsets:

```
library(leaps)
best <- regsubsets(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 +
                     X7 + X8, data = surg, nvmax = 8)
s <- summary(best)
s$which         # the predictors of the best model of each size
s$adjr2; s$cp   # their adjusted R2 and Cp
```

# In R: the best model of each size

The package `leaps` searches all the subsets:

```
library(leaps)
best <- regsubsets(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 +
                     X7 + X8, data = surg, nvmax = 8)
s <- summary(best)
s$which         # the predictors of the best model of each size
s$adjr2; s$cp   # their adjusted R2 and Cp
```

- `nbest = 3` keeps the three best models of each size.
- `s$bic` is our $BIC_p$ up to a constant: same ranking.

# In R: the best model of each size

The package `leaps` searches all the subsets:

```
library(leaps)
best <- regsubsets(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 +
                     X7 + X8, data = surg, nvmax = 8)
s <- summary(best)
s$which         # the predictors of the best model of each size
s$adjr2; s$cp   # their adjusted R2 and Cp
```

- `nbest = 3` keeps the three best models of each size.
- `s$bic` is our $BIC_p$ up to a constant: same ranking.

⚠ `regsubsets` sees columns, not variables. It may keep one indicator of a factor without the other (here $X_8$ without $X_7$): allowed, but $X_8$ then compares severe use with none and moderate together. It may also keep a square or a product without its lower-order terms, which we do not allow (Lectures 16 and 17): add those terms back before using such a model.

# What the criteria do not do

- They rank models, they do not check them. The residual analysis of the chosen model is still to be done, and so is its validation.
- They often leave several models that are nearly as good. Choose among them with what you know of the subject (which predictors should matter, which are easy to measure), prefer the simpler one, and say how you chose.

# What the criteria do not do

- They rank models, they do not check them. The residual analysis of the chosen model is still to be done, and so is its validation.
- They often leave several models that are nearly as good. Choose among them with what you know of the subject (which predictors should matter, which are easy to measure), prefer the simpler one, and say how you chose.
- It uses the same response and the same cases. With missing values, R drops different rows for different models: remove the incomplete rows first.
- The model that wins was picked because it fits *these* data well: its $p$-values and its $R^2$ are somewhat optimistic.

# What the criteria do not do

- They rank models, they do not check them. The residual analysis of the chosen model is still to be done, and so is its validation.
- They often leave several models that are nearly as good. Choose among them with what you know of the subject (which predictors should matter, which are easy to measure), prefer the simpler one, and say how you chose.
- It uses the same response and the same cases. With missing values, R drops different rows for different models: remove the incomplete rows first.
- The model that wins was picked because it fits *these* data well: its $p$-values and its $R^2$ are somewhat optimistic.

In a report, "the software chose this model" is not a justification: give the criterion, the candidates, and the reason for the final choice!

# Automatic search procedures

# When there are too many models

With 30 candidate predictors there are $2^{30}$ models, about a billion. A search procedure moves through them one predictor at a time:

- Backward elimination: start from the model with all the predictors, then remove the predictor whose removal improves the criterion most and stop when no removal improves it.
- Forward selection: start from the intercept alone, then add the predictor that improves the criterion most and stop when no addition improves it.
- Stepwise regression: forward, but after each addition, check whether a predictor already in can now be removed.

# When there are too many models

With 30 candidate predictors there are $2^{30}$ models, about a billion. A search procedure moves through them one predictor at a time:

- Backward elimination: start from the model with all the predictors, then remove the predictor whose removal improves the criterion most and stop when no removal improves it.
- Forward selection: start from the intercept alone, then add the predictor that improves the criterion most and stop when no addition improves it.
- Stepwise regression: forward, but after each addition, check whether a predictor already in can now be removed.

The criterion can be a test (enter or stay if the $p$-value of the $t$ test is below a threshold, often 0.10 to 0.20). R's `step()` uses the AIC.

# Backward elimination with `step()`

```
full <- lm(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 +
             X7 + X8, data = surg)
back <- step(full, direction = "backward", trace = 0)
back$anova
```

```
  Step Df    Deviance Resid. Df Resid. Dev       AIC
1      NA          NA        45   1.970510 -160.7773
2 - X4  1 0.001262737        46   1.971773 -162.7428
3 - X7  1 0.032562516        47   2.004335 -163.8583
```

# Backward elimination with `step()`

```
full <- lm(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 +
             X7 + X8, data = surg)
back <- step(full, direction = "backward", trace = 0)
back$anova
```

```
  Step Df    Deviance Resid. Df Resid. Dev       AIC
1      NA          NA        45   1.970510 -160.7773
2 - X4  1 0.001262737        46   1.971773 -162.7428
3 - X7  1 0.032562516        47   2.004335 -163.8583
```

`trace = 0` turns off the table that `step()` prints at each step; `back$anova` keeps the path. `Resid. Dev` is $SSE_p$. $X_4$ goes first, then $X_7$; any other removal would raise the AIC. The search stops at $X_1, X_2, X_3, X_5, X_6, X_8$: the smallest AIC of the table.

# Backward elimination with `step()`

```
full <- lm(log(Y) ~ X1 + X2 + X3 + X4 + X5 + X6 +
             X7 + X8, data = surg)
back <- step(full, direction = "backward", trace = 0)
back$anova
```

```
  Step Df    Deviance Resid. Df Resid. Dev       AIC
1      NA          NA        45   1.970510 -160.7773
2 - X4  1 0.001262737        46   1.971773 -162.7428
3 - X7  1 0.032562516        47   2.004335 -163.8583
```

`trace = 0` turns off the table that `step()` prints at each step; `back$anova` keeps the path. `Resid. Dev` is $SSE_p$. $X_4$ goes first, then $X_7$; any other removal would raise the AIC. The search stops at $X_1, X_2, X_3, X_5, X_6, X_8$: the smallest AIC of the table.

With `k = log(n)` in `step()`, the criterion is the BIC: two more predictors go, and it stops at $X_1, X_2, X_3, X_8$.

# Forward selection and stepwise regression

```
null <- lm(log(Y) ~ 1, data = surg)
forw <- step(null, scope = formula(full),
             direction = "forward", trace = 0)
```

# Forward selection and stepwise regression

```
null <- lm(log(Y) ~ 1, data = surg)
forw <- step(null, scope = formula(full),
             direction = "forward", trace = 0)
```

Forward selection adds $X_3, X_2, X_8, X_1, X_6, X_5$, in this order, and stops at the same model. With `direction = "both"`, the stepwise search takes the same path here.

# Forward selection and stepwise regression

```
null <- lm(log(Y) ~ 1, data = surg)
forw <- step(null, scope = formula(full),
             direction = "forward", trace = 0)
```

Forward selection adds $X_3, X_2, X_8, X_1, X_6, X_5$, in this order, and stops at the same model. With `direction = "both"`, the stepwise search takes the same path here.

- A search follows one path: it can miss the best model, and the three procedures can disagree.
- A `factor` column enters or leaves with all its indicators. Here $X_7$ and $X_8$ are two separate columns, so $X_7$ could go alone: the model then compares severe use with the rest.
- It is a way to find candidates, not a way to choose for you.

# Summary

- A response whose spread grows with its level, with skewed errors: try $\ln Y$. One more unit of $X_k$ then multiplies the median of $Y$ by $e^{b_k}$.
- A curve in the residuals against a predictor, with well-behaved errors: transform the predictor. With $\ln X$, the coefficient is read for a change of 1% in $X$.
- $R^2_p$ cannot compare models of different sizes. Adjusted $R^2$, $C_p$, AIC, BIC and PRESS each penalize more parameters.
- $C_p$ estimates the total mean squared error: bias from the predictors left out, plus $p$.
- The criteria rank models for the same response on the same cases. They do not replace the residual analysis, the context, or validation.