# STAC67: Regression Analysis

## Lecture 20

Thibault Randrianarisoa

November 20, 2026

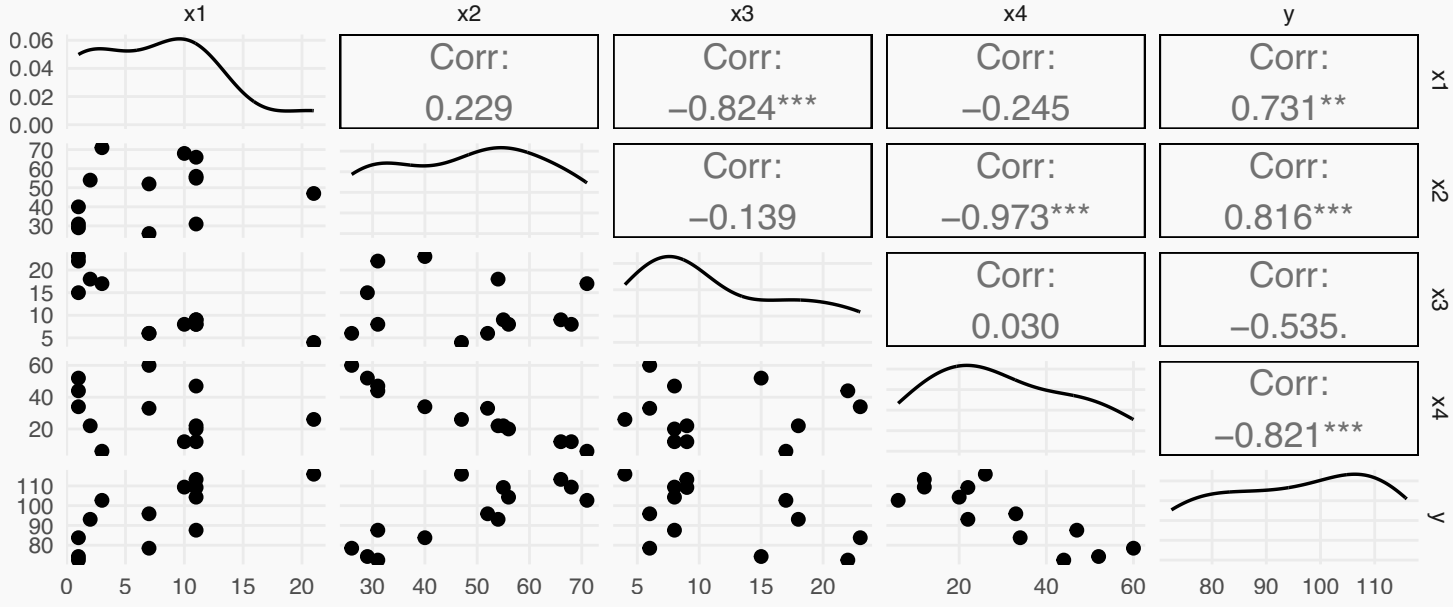
University of Toronto Scarborough

# Automatic search procedures

# Step by step: the cement data

13 batches of cement: the heat released as it sets ($y$, calories per gram), and the percentages of four ingredients ($x_1$ to $x_4$), which add up to 95 to 99% of each batch.

```
cement <- MASS::cement
GGally::ggpairs(cement)
```



# Searching by BIC

The predictors are strongly related: $r(x_2, x_4) = -0.97$ and $r(x_1, x_3) = -0.82$.

# Searching by BIC

The predictors are strongly related: $r(x_2, x_4) = -0.97$ and $r(x_1, x_3) = -0.82$.

Criterion: $BIC_p = n \ln(SSE_p/n) + p \ln n$, with $n = 13$ and $\ln 13 = 2.56$. Small is good. At each step, make the move with the smallest BIC, then stop when no move lowers it.

# Searching by BIC

The predictors are strongly related: $r(x_2, x_4) = -0.97$ and $r(x_1, x_3) = -0.82$.

Criterion: $BIC_p = n \ln(SSE_p/n) + p \ln n$, with $n = 13$ and $\ln 13 = 2.56$. Small is good. At each step, make the move with the smallest BIC, then stop when no move lowers it.

`ggpairs`, from the package `GGally`, draws every pair of variables: the scatterplots below the diagonal, the correlations above it (with stars for their tests), and the distribution of each variable on the diagonal.

# Forward selection, step by step

Start from the intercept; at each step, try adding each predictor that is not in yet.

| Step | 1 |
|---|---|
| Current model | intercept |
| add $x_1$ | 64.65 |
| add $x_2$ | 60.31 |
| add $x_3$ | 70.20 |
| add $x_4$ | 59.98 |
| no change | 72.01 |

# Forward selection, step by step

Start from the intercept; at each step, try adding each predictor that is not in yet.

| Step | 1 | 2 |
|---|---|---|
| Current model | intercept | $x_4$ |
| add $x_1$ | 64.65 | 30.44 |
| add $x_2$ | 60.31 | 62.32 |
| add $x_3$ | 70.20 | 41.55 |
| add $x_4$ | 59.98 | in |
| no change | 72.01 | 59.98 |

# Forward selection, step by step

Start from the intercept; at each step, try adding each predictor that is not in yet.

| Step | 1 | 2 | 3 |
|---|---|---|---|
| Current model | intercept | $x_4$ | $x_1, x_4$ |
| add $x_1$ | 64.65 | 30.44 | in |
| add $x_2$ | 60.31 | 62.32 | 27.23 |
| add $x_3$ | 70.20 | 41.55 | 27.99 |
| add $x_4$ | 59.98 | in | in |
| no change | 72.01 | 59.98 | 30.44 |

# Forward selection, step by step

Start from the intercept; at each step, try adding each predictor that is not in yet.

| Step | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Current model | intercept | $x_4$ | $x_1, x_4$ | $x_1, x_2, x_4$ |
| add $x_1$ | 64.65 | 30.44 | in | in |
| add $x_2$ | 60.31 | 62.32 | 27.23 | in |
| add $x_3$ | 70.20 | 41.55 | 27.99 | 29.77 |
| add $x_4$ | 59.98 | in | in | in |
| no change | 72.01 | 59.98 | 30.44 | 27.23 |

# Forward selection, step by step

Start from the intercept; at each step, try adding each predictor that is not in yet.

| Step | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Current model | intercept | $x_4$ | $x_1, x_4$ | $x_1, x_2, x_4$ |
| add $x_1$ | 64.65 | 30.44 | in | in |
| add $x_2$ | 60.31 | 62.32 | 27.23 | in |
| add $x_3$ | 70.20 | 41.55 | 27.99 | 29.77 |
| add $x_4$ | 59.98 | in | in | in |
| no change | 72.01 | 59.98 | 30.44 | 27.23 |

Forward selection adds $x_4$, then $x_1$, then $x_2$, and stops: adding $x_3$ would raise the BIC. Final model: $x_1, x_2, x_4$.

# Backward elimination, step by step

Start from the four predictors; at each step, try removing each predictor still in.

| Step | 1 |
|---|---|
| Current model | $x_1, x_2, x_3, x_4$ |
| remove $x_1$ | 32.84 |
| remove $x_2$ | 27.99 |
| remove $x_3$ | 27.23 |
| remove $x_4$ | 27.27 |
| no change | 29.77 |

# Backward elimination, step by step

Start from the four predictors; at each step, try removing each predictor still in.

| Step | 1 | 2 |
|---|---|---|
| Current model | $x_1, x_2, x_3, x_4$ | $x_1, x_2, x_4$ |
| remove $x_1$ | 32.84 | 62.32 |
| remove $x_2$ | 27.99 | 30.44 |
| remove $x_3$ | 27.23 | out |
| remove $x_4$ | 27.27 | 27.11 |
| no change | 29.77 | 27.23 |

# Backward elimination, step by step

Start from the four predictors; at each step, try removing each predictor still in.

| Step | 1 | 2 | 3 |
|---|---|---|---|
| Current model | $x_1, x_2, x_3, x_4$ | $x_1, x_2, x_4$ | $x_1, x_2$ |
| remove $x_1$ | 32.84 | 62.32 | 60.31 |
| remove $x_2$ | 27.99 | 30.44 | 64.65 |
| remove $x_3$ | 27.23 | out | out |
| remove $x_4$ | 27.27 | 27.11 | out |
| no change | 29.77 | 27.23 | 27.11 |

# Backward elimination, step by step

Start from the four predictors; at each step, try removing each predictor still in.

| Step | 1 | 2 | 3 |
|---|---|---|---|
| Current model | $x_1, x_2, x_3, x_4$ | $x_1, x_2, x_4$ | $x_1, x_2$ |
| remove $x_1$ | 32.84 | 62.32 | 60.31 |
| remove $x_2$ | 27.99 | 30.44 | 64.65 |
| remove $x_3$ | 27.23 | out | out |
| remove $x_4$ | 27.27 | 27.11 | out |
| no change | 29.77 | 27.23 | 27.11 |

Backward elimination removes $x_3$, then $x_4$, and stops. Final model: $x_1, x_2$, not the same as forward selection.

With the AIC (the default in `step()`), the three searches all stop at $x_1, x_2, x_4$: the criterion matters too.

# Stepwise regression: a predictor can leave

Forward as before: $x_4$, then $x_1$, then $x_2$. But after each addition, removals are tried too:

| Step | 4 |
|---|---|
| Current model | $x_1, x_2, x_4$ |
| add $x_3$ | 29.77 |
| add $x_4$ | |
| remove $x_1$ | 62.32 |
| remove $x_2$ | 30.44 |
| remove $x_4$ | 27.11 |
| no change | 27.23 |

# Stepwise regression: a predictor can leave

Forward as before: $x_4$, then $x_1$, then $x_2$. But after each addition, removals are tried too:

| Step | 4 | 5 |
|---|---|---|
| Current model | $x_1, x_2, x_4$ | $x_1, x_2$ |
| add $x_3$ | 29.77 | 27.27 |
| add $x_4$ | | 27.23 |
| remove $x_1$ | 62.32 | 60.31 |
| remove $x_2$ | 30.44 | 64.65 |
| remove $x_4$ | 27.11 | |
| no change | 27.23 | 27.11 |

# Stepwise regression: a predictor can leave

Forward as before: $x_4$, then $x_1$, then $x_2$. But after each addition, removals are tried too:

| Step | 4 | 5 |
|---|---|---|
| Current model | $x_1, x_2, x_4$ | $x_1, x_2$ |
| add $x_3$ | 29.77 | 27.27 |
| add $x_4$ | | 27.23 |
| remove $x_1$ | 62.32 | 60.31 |
| remove $x_2$ | 30.44 | 64.65 |
| remove $x_4$ | 27.11 | |
| no change | 27.23 | 27.11 |

$x_4$ was the best single predictor, but once $x_2$ is in ($r = -0.97$), it adds almost nothing: it leaves. Stepwise ends at $x_1, x_2$, like backward elimination. Forward selection cannot undo a choice.

# In R: `step()` prints every step

```
null <- lm(y ~ 1, data = cement)
step(null, scope = ~ x1 + x2 + x3 + x4,
     direction = "both", k = log(13))
```

# In R: `step()` prints every step

```r
null <- lm(y ~ 1, data = cement)
step(null, scope = ~ x1 + x2 + x3 + x4,
     direction = "both", k = log(13))
```

The step where $x_4$ leaves:

```
Step:  AIC=27.23
y ~ x4 + x1 + x2

       Df Sum of Sq    RSS    AIC
- x4    1      9.93  57.90 27.115
<none>               47.97 27.234
+ x3    1      0.11  47.86 29.769
- x2    1     26.79  74.76 30.437
- x1    1    820.91 868.88 62.324
```

# In R: `step()` prints every step

```r
null <- lm(y ~ 1, data = cement)
step(null, scope = ~ x1 + x2 + x3 + x4,
     direction = "both", k = log(13))
```

The step where $x_4$ leaves:

```
Step:  AIC=27.23
y ~ x4 + x1 + x2

       Df Sum of Sq    RSS    AIC
- x4    1      9.93  57.90 27.115
<none>               47.97 27.234
+ x3    1      0.11  47.86 29.769
- x2    1     26.79  74.76 30.437
- x1    1    820.91 868.88 62.324
```

Each block is one step: the current model, then every move, sorted. The column says AIC, but with `k = log(13)` it is the BIC.

# The surgical unit: 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
```

# The surgical unit: 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, but `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.

# The surgical unit: 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, but `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 $X_1, X_2, X_3, X_8$.

# The surgical unit: forward selection and stepwise

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

# The surgical unit: forward selection and stepwise

```
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.

# The surgical unit: forward selection and stepwise

```
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.

# Model validation

# Why validate?

The chosen model was picked because it fits *these* data well: its $MSE$ and its $R^2$ flatter it.

Its $t$ tests too. A $t$ test assumes the model was fixed before seeing the data. Here the predictors were kept *because* their $t$ statistics were large on these data (partly by chance): tested again on the same data, they look more significant than they are, and the $p$-values are too small.

# Why validate?

The chosen model was picked because it fits *these* data well: its $MSE$ and its $R^2$ flatter it.

Its $t$ tests too. A $t$ test assumes the model was fixed before seeing the data. Here the predictors were kept *because* their $t$ statistics were large on these data (partly by chance): tested again on the same data, they look more significant than they are, and the $p$-values are too small.

We need a way to confirm that the final model is sound and effective for the purpose it was intended for.

Validation: check how the model does on data that were used neither to choose it nor to fit it.

- Collect new data, and predict them.
- Compare the coefficients with theory, or with earlier studies.
- Split the data: a training set to choose and fit the model, and a validation set (or test set), kept aside from the start.

Potentially, the fit on the dataset used for model selection is wrong

70%

30%

# Why validate?

The chosen model was picked because it fits *these* data well: its $MSE$ and its $R^2$ flatter it.

Its $t$ tests too. A $t$ test assumes the model was fixed before seeing the data. Here the predictors were kept *because* their $t$ statistics were large on these data (partly by chance): tested again on the same data, they look more significant than they are, and the $p$-values are too small.

We need a way to confirm that the final model is sound and effective for the purpose it was intended for.

Validation: check how the model does on data that were used neither to choose it nor to fit it.

- Collect new data, and predict them.
- Compare the coefficients with theory, or with earlier studies.
- Split the data: a training set to choose and fit the model, and a validation set (or test set), kept aside from the start.

The surgical unit: 108 patients, split at random. The 54 of Lecture 18 are the training set; the other 54 were kept aside.

# The mean squared prediction error

Fit the model on the training set. For each of the $n^*$ cases of the validation set, predict $Y_i$ with the training coefficients:

$$MSPE = \frac{1}{n^*} \sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2.$$

$Y_i$: $i$th obs. of validation set

$\hat{Y}_i$: $i$th fitted value of validation set from model fitted on the training set

# The mean squared prediction error

Fit the model on the training set. For each of the $n^*$ cases of the validation set, predict $Y_i$ with the training coefficients:

$$MSPE = \frac{1}{n^*} \sum_{i=1}^{n^*} \left(Y_i - \hat{Y}_i\right)^2.$$

- Divide by $n^*$, not $n^* - p$: no parameter is estimated on the validation set.
- Compare with the $MSE$ of the training fit. If the $MSPE$ is close to it, the $MSE$ is a fair measure of how well the model predicts. If it is much larger, the model fits the training set better than it predicts new cases.

# The mean squared prediction error

Fit the model on the training set. For each of the $n^*$ cases of the validation set, predict $Y_i$ with the training coefficients:

$$MSPE = \frac{1}{n^*} \sum_{i=1}^{n^*} \left(Y_i - \hat{Y}_i\right)^2.$$

- Divide by $n^*$, not $n^* - p$: no parameter is estimated on the validation set.
- Compare with the $MSE$ of the training fit. If the $MSPE$ is close to it, the $MSE$ is a fair measure of how well the model predicts. If it is much larger, the model fits the training set better than it predicts new cases.

Once the model is validated, refit it on all the data for the final estimates.



# Exercise: the 54 patients kept aside

The model with $X_1, X_2, X_3, X_8$ (chosen by BIC and PRESS) is fitted on the 54 training patients: $SSE = 2.1778$, with $p = 5$.

It then predicts the 54 patients kept aside: $\sum_{i=1}^{54}(Y_i - \hat{Y}_i)^2 = 4.174$.

1. Compute the $MSE$ of the training fit, and the $MSPE$.
2. Does the model predict new patients as well as its $MSE$ suggests?

# Exercise: solution, and the three candidates

$MSE = 2.1778/49 = 0.0444$ and $MSPE = 4.174/54 = 0.0773$. The same for Wednesday's other two candidates, $MSE$ on the training set and $MSPE$ on the validation set:

| Predictors | Chosen by | $p$ | $MSE$ | $MSPE$ |
|---|---|---|---|---|
| $X_1, X_2, X_3, X_8$ | BIC, PRESS | 5 | 0.0444 | 0.0773 |
| $X_1, X_2, X_3, X_6, X_8$ | $C_p$ | 6 | 0.0434 | 0.0764 |
| $X_1, X_2, X_3, X_5, X_6, X_8$ | $R^2_a$, AIC | 7 | 0.0426 | 0.0794 |

$\sim 10^{-2}$

# Exercise: solution, and the three candidates

$MSE = 2.1778/49 = 0.0444$ and $MSPE = 4.174/54 = 0.0773$. The same for Wednesday's other two candidates, $MSE$ on the training set and $MSPE$ on the validation set:

| Predictors | Chosen by | $p$ | $MSE$ | $MSPE$ |
|---|---|---|---|---|
| $X_1, X_2, X_3, X_8$ | BIC, PRESS | 5 | 0.0444 | 0.0773 |
| $X_1, X_2, X_3, X_6, X_8$ | $C_p$ | 6 | 0.0434 | 0.0764 |
| $X_1, X_2, X_3, X_5, X_6, X_8$ | $R^2_a$, AIC | 7 | 0.0426 | 0.0794 |

- The $MSPE$ is 1.7 to 1.9 times the $MSE$: predicting new patients is harder than fitting the old ones. It stays of the same order: the models are acceptable, with a prediction error near 0.08, not 0.04.
- The three predict equally well: the extra predictors buy nothing. Keep $X_1, X_2, X_3, X_8$.

# Exercise: solution, and the three candidates

$MSE = 2.1778/49 = 0.0444$ and $MSPE = 4.174/54 = 0.0773$. The same for Wednesday's other two candidates, $MSE$ on the training set and $MSPE$ on the validation set:

| Predictors | Chosen by | $p$ | $MSE$ | $MSPE$ |
|---|---|---|---|---|
| $X_1, X_2, X_3, X_8$ | BIC, PRESS | 5 | 0.0444 | 0.0773 |
| $X_1, X_2, X_3, X_6, X_8$ | $C_p$ | 6 | 0.0434 | 0.0764 |
| $X_1, X_2, X_3, X_5, X_6, X_8$ | $R^2_a$, AIC | 7 | 0.0426 | 0.0794 |

- The $MSPE$ is 1.7 to 1.9 times the $MSE$: predicting new patients is harder than fitting the old ones. It stays of the same order: the models are acceptable, with a prediction error near 0.08, not 0.04.
- The three predict equally well: the extra predictors buy nothing. Keep $X_1, X_2, X_3, X_8$.

Refitted on the new patients, the coefficient of age $X_5$ changes sign: one more reason to leave it out.

# After validation

If we refit the same model on the validation set, and compare with the training fit:

- $MSPE$ close to the refitted $MSE$ (here 0.0773 against 0.0775): the estimates from the training set predict the new cases as well as estimates computed from the new cases themselves. Only the training $MSE$ was optimistic. Report the $MSPE$ as its prediction error.

# Using the validation set well

- Choose among models on the training set only, with the criteria.
- The validation set checks the final candidate, once. Each extra use to choose makes its $MSPE$ more optimistic: it becomes part of the training.

# Using the validation set well

- Choose among models on the training set only, with the criteria.
- The validation set checks the final candidate, once. Each extra use to choose makes its $MSPE$ more optimistic: it becomes part of the training.
- The model is validated if, on the validation set: its $MSPE$ is close to the training $MSE$; refitted there, its coefficients keep their signs and similar sizes; its residuals show no new pattern. There is no fixed rule, it is somewhat subjective.

# Using the validation set well

- Choose among models on the training set only, with the criteria.
- The validation set checks the final candidate, once. Each extra use to choose makes its $MSPE$ more optimistic: it becomes part of the training.
- The model is validated if, on the validation set: its $MSPE$ is close to the training $MSE$; refitted there, its coefficients keep their signs and similar sizes; its residuals show no new pattern. There is no fixed rule, it is somewhat subjective.
- Once the model is validated, refit it on all the data (here 108 patients) for the final estimates.

# A validation set in R

The acceleration data (390 cars) and Lecture 18's three models for `acc`: 273 cars to train, 117 to validate.

```
set.seed(1)
train <- sample(nrow(autompg), 273)
fit <- lm(acc ~ wt + log(hp), data = autompg[train, ])
pred <- predict(fit, newdata = autompg[-train, ])
mean((autompg$acc[-train] - pred)^2)          # MSPE
```

# A validation set in R

The acceleration data (390 cars) and Lecture 18's three models for `acc`: 273 cars to train, 117 to validate.

```
set.seed(1)
train <- sample(nrow(autompg), 273)
fit <- lm(acc ~ wt + log(hp), data = autompg[train, ])
pred <- predict(fit, newdata = autompg[-train, ])
mean((autompg$acc[-train] - pred)^2)          # MSPE
```

| | set.seed(1) | | set.seed(2) | |
|---|---|---|---|---|
| Model | $MSE$ | $MSPE$ | $MSE$ | $MSPE$ |
| `wt + hp` | 3.33 | 2.50 | 2.90 | 3.49 |
| `wt + log(hp)` | 2.92 | 2.15 | 2.68 | 2.78 |
| `wt + hp + I(hp^2)` | 2.84 | 2.13 | 2.59 | 2.74 |

# A validation set in R

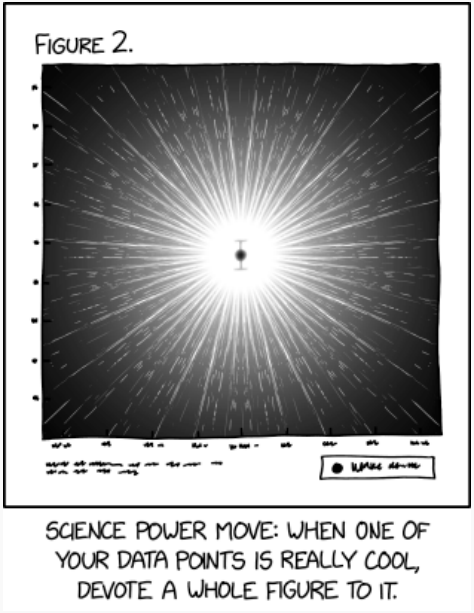
The acceleration data (390 cars) and Lecture 18's three models for `acc`: 273 cars to train, 117 to validate.

```
set.seed(1)
train <- sample(nrow(autompg), 273)
fit <- lm(acc ~ wt + log(hp), data = autompg[train, ])
pred <- predict(fit, newdata = autompg[-train, ])
mean((autompg$acc[-train] - pred)^2)          # MSPE
```

| | set.seed(1) | | set.seed(2) | |
|---|---|---|---|---|
| Model | $MSE$ | $MSPE$ | $MSE$ | $MSPE$ |
| `wt + hp` | 3.33 | 2.50 | 2.90 | 3.49 |
| `wt + log(hp)` | 2.92 | 2.15 | 2.68 | 2.78 |
| `wt + hp + I(hp^2)` | 2.84 | 2.13 | 2.59 | 2.74 |

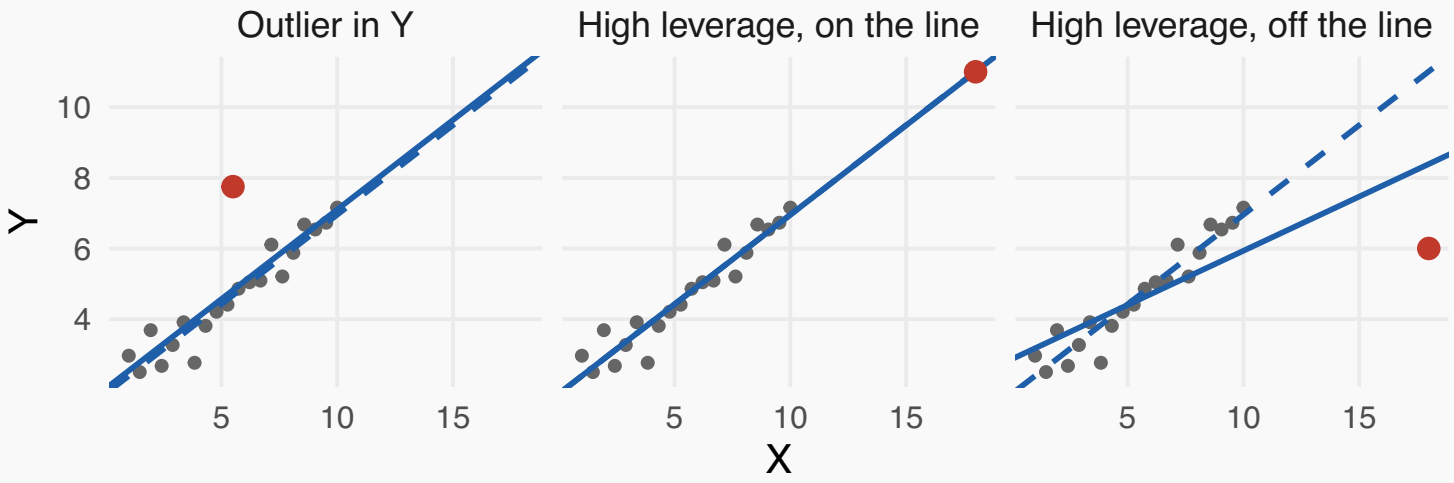
The ranking holds, but the $MSPE$ depends on which cars fell in the validation set: below the $MSE$ with one split, above it with the other. The $MSPE$ is itself random: it depends on the split.

# Outlying and influential cases



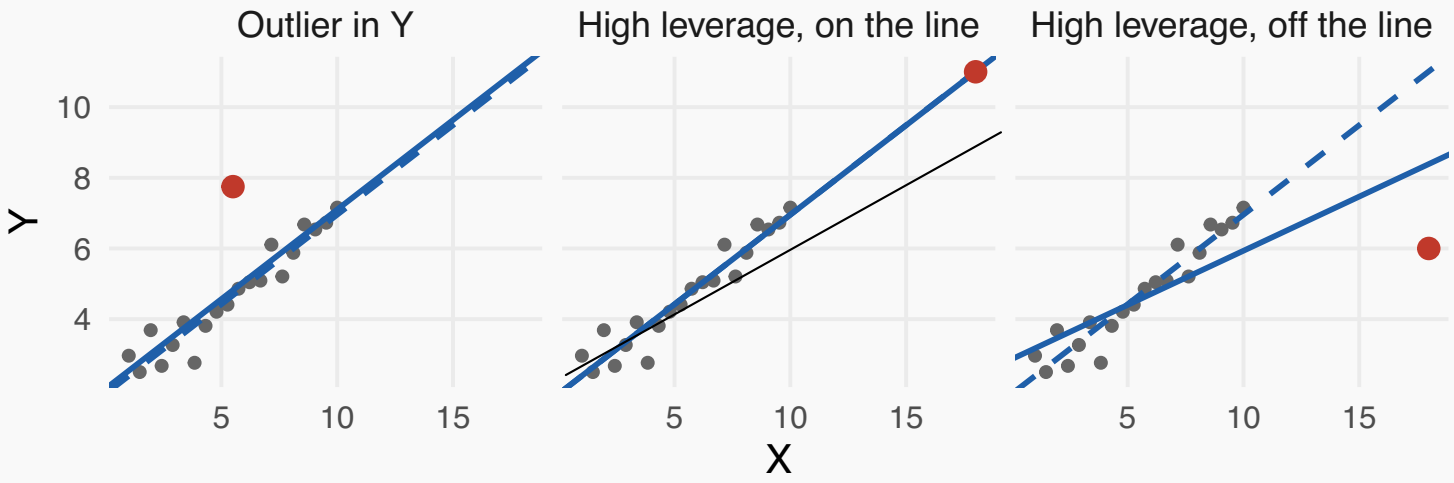

xkcd.com/2713, *Data Point* (CC BY-NC 2.5)

# Three kinds of unusual cases




Solid: the fit with all the points. Dashed: the fit without the red one.

# Three kinds of unusual cases




Solid: the fit with all the points. Dashed: the fit without the red one.

- Outlier in $Y$: far from the fitted line, a large residual (left).
- Outlier in $X$: far from the other cases in the space of the predictors, a high leverage (middle and right).
- Influential: removing it changes the fit noticeably. That usually takes both: far in $X$ and off the line of the other points (right).

# Residuals, scaled

From Lecture 12: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $\boldsymbol{\sigma}^2\{\boldsymbol{e}\} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, so

$$Var(e_i) = (1 - h_{ii})\,\sigma^2.$$

# Residuals, scaled

From Lecture 12: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $\boldsymbol{\sigma}^2\{\boldsymbol{e}\} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, so

$$Var(e_i) = (1 - h_{ii})\,\sigma^2.$$

A case with a high leverage pulls the fit towards itself: its residual is small even when the case is unusual. Divide each residual by its own standard error, the studentized residual:

$$r_i = \frac{e_i}{\sqrt{MSE\,(1 - h_{ii})}}.$$

$\Rightarrow \dfrac{e_i}{SE(e_i)}$

# Residuals, scaled

From Lecture 12: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $\boldsymbol{\sigma}^2\{\boldsymbol{e}\} = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, so

$$Var(e_i) = (1 - h_{ii})\,\sigma^2.$$

A case with a high leverage pulls the fit towards itself: its residual is small even when the case is unusual. Divide each residual by its own standard error, the studentized residual:

$$r_i = \frac{e_i}{\sqrt{MSE\,(1 - h_{ii})}}.$$

$e_i/\sqrt{MSE}$, the semistudentized residual, ignores the leverage.

# Deleted residuals

Better: predict case $i$ with a fit that did not use it, $\hat{Y}_{i(i)}$. The deleted residual is (PRESS, Lecture 18)

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}.$$

# Deleted residuals

$$Var(d_i) = \frac{1}{(1-h_{ii})^2} Var(e_i) = \frac{\sigma^2}{1-h_{ii}}$$

Better: predict case $i$ with a fit that did not use it, $\hat{Y}_{i(i)}$. The deleted residual is (PRESS, Lecture 18)

predicted-value from model without i in the training set

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}.$$

Divided by its standard error, the studentized deleted residual:

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}\,(1-h_{ii})}} = e_i \left[\frac{n-p-1}{SSE\,(1-h_{ii}) - e_i^2}\right]^{1/2} \sim t(n-p-1),$$

with $MSE_{(i)}$ the $MSE$ of the fit without case $i$.

# Deleted residuals

Better: predict case $i$ with a fit that did not use it, $\hat{Y}_{i(i)}$. The deleted residual is (PRESS, Lecture 18)

$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}.$$

Divided by its standard error, the studentized deleted residual:

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}\,(1 - h_{ii})}} = e_i \left[\frac{n - p - 1}{SSE\,(1 - h_{ii}) - e_i^2}\right]^{1/2} \sim t(n - p - 1),$$

with $MSE_{(i)}$ the $MSE$ of the fit without case $i$.

One fit is enough: no need to refit $n$ times. In R: `rstudent(fit)`.

# Is a case an outlier in $Y$?

An outlier has a large $|t_i|$. One test per case, so $n$ tests: with the Bonferroni correction, case $i$ is an outlier if

$$|t_i| > t\big(1 - \tfrac{\alpha}{2n};\ n - p - 1\big).$$

# Is a case an outlier in $Y$?

An outlier has a large $|t_i|$. One test per case, so $n$ tests: with the Bonferroni correction, case $i$ is an outlier if

$$|t_i| > t\big(1 - \tfrac{\alpha}{2n};\ n - p - 1\big).$$

The surgical unit, with the model kept on Wednesday ($X_1, X_2, X_3, X_8$; $n = 54$, $p = 5$): $t(1 - 0.05/108;\ 48) = 3.53$.

The largest $|t_i|$ is 3.36, for case 17: no outlier at the 5% level.

# Is a case an outlier in $Y$?

An outlier has a large $|t_i|$. One test per case, so $n$ tests: with the Bonferroni correction, case $i$ is an outlier if

$$|t_i| > t\big(1 - \tfrac{\alpha}{2n};\ n - p - 1\big).$$

The surgical unit, with the model kept on Wednesday ($X_1, X_2, X_3, X_8$; $n = 54$, $p = 5$): $t(1 - 0.05/108;\ 48) = 3.53$.

The largest $|t_i|$ is 3.36, for case 17: no outlier at the 5% level.

Without the correction, about 5% of the cases would cross $t(0.975;\ 48) = 2.01$ by chance alone: two or three of 54.

# An outlier in $X$: leverage

$h_{ii} = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i$, the leverage ofobservation $i$, measures how far case $i$ is from the centre of the predictors (Lectures 11 and 12).

$0 \leq h_{ii} \leq 1$ and $\sum_i h_{ii} = p$: on average $p/n$.

# An outlier in $X$: leverage

$h_{ii} = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i$, the leverage ofobservation $i$, measures how far case $i$ is from the centre of the predictors (Lectures 11 and 12).

$0 \leq h_{ii} \leq 1$ and $\sum_i h_{ii} = p$: on average $p/n$.

- High leverage: $h_{ii} > 2p/n$. Very high: $h_{ii} > 0.5$.
- The surgical unit: $2p/n = 0.185$. Six patients are above, the highest at 0.31 (case 38); none above 0.5.

# An outlier in $X$: leverage

$h_{ii} = \boldsymbol{X}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i$, the leverage ofobservation $i$, measures how far case $i$ is from the centre of the predictors (Lectures 11 and 12).

$0 \leq h_{ii} \leq 1$ and $\sum_i h_{ii} = p$: on average $p/n$.

- High leverage: $h_{ii} > 2p/n$. Very high: $h_{ii} > 0.5$.
- The surgical unit: $2p/n = 0.185$. Six patients are above, the highest at 0.31 (case 38); none above 0.5.

Hidden extrapolation: a new case can be within the range of each predictor and still far from the data. Compare $h_{new} = \boldsymbol{X}_{new}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_{new}$ with the $h_{ii}$ of the data.

# Influence on the fitted values

Remove case $i$, refit, and measure quantitatively what moves, in standard errors.

DFFITS: the change in its own fitted value,

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}\, h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

Influential if $|DFFITS_i| > 1$ (small or medium $n$), or $2\sqrt{p/n}$ (large $n$).

# Influence on the fitted values

Remove case $i$, refit, and measure quantitatively what moves, in standard errors.

DFFITS: the change in its own fitted value,

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}\, h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}.$$

Influential if $|DFFITS_i| > 1$ (small or medium $n$), or $2\sqrt{p/n}$ (large $n$).

Cook's distance: the change in all $n$ fitted values,

$$D_i = \frac{\sum_j \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p\, MSE} = \frac{e_i^2}{p\, MSE} \frac{h_{ii}}{(1 - h_{ii})^2}.$$

Compare with $F(p, n - p)$: below its 10th or 20th percentile, little influence; near its median or above, major influence.

# Influence on the coefficients: DFBETAS

How many standard errors $\hat{\beta}_k$ moves without case $i$:

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}\, c_{kk}}},$$

with $c_{kk}$ the diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ for $\hat{\beta}_k$.

# Influence on the coefficients: DFBETAS

How many standard errors $\hat{\beta}_k$ moves without case $i$:

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}\, c_{kk}}},$$

with $c_{kk}$ the diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ for $\hat{\beta}_k$.

Influential if $|DFBETAS_{k(i)}| > 1$ (small or medium $n$), or $2/\sqrt{n}$ (large $n$).

# Influence on the coefficients: DFBETAS

How many standard errors $\hat{\beta}_k$ moves without case $i$:

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}\, c_{kk}}},$$

with $c_{kk}$ the diagonal entry of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ for $\hat{\beta}_k$.

Influential if $|DFBETAS_{k(i)}| > 1$ (small or medium $n$), or $2/\sqrt{n}$ (large $n$).

The surgical unit: one value above 1, case 17 on $\hat{\beta}_3$ (enzyme score), $-1.14$. Without case 17, $\hat{\beta}_3$ goes from 0.0154 to 0.0169; the other coefficients barely move.

# Exercise: four patients

The surgical unit, model $X_1, X_2, X_3, X_8$ for $\ln Y$: $n = 54$, $p = 5$, $MSE = 0.0444$.

| Case | $e_i$ | $h_{ii}$ | $t_i$ | $DFFITS_i$ | $D_i$ |
|---|---|---|---|---|---|
| 1 | $0.069$ | 0.035 | $0.33$ | | |
| 17 | $0.594$ | 0.150 | $3.36$ | | |
| 38 | $-0.228$ | 0.306 | $-1.31$ | | |
| 45 | $-0.453$ | 0.083 | $-2.35$ | | |

1. Compute $DFFITS_i$ and $D_i$ for the four patients.
2. Which are outliers in $Y$ ($t(1 - 0.05/108;\ 48) = 3.53$), outliers in $X$ ($2p/n = 0.185$), influential?

The 10th, 20th and 50th percentiles of $F(5, 49)$ are 0.318, 0.466 and 0.882.

# Exercise: solution

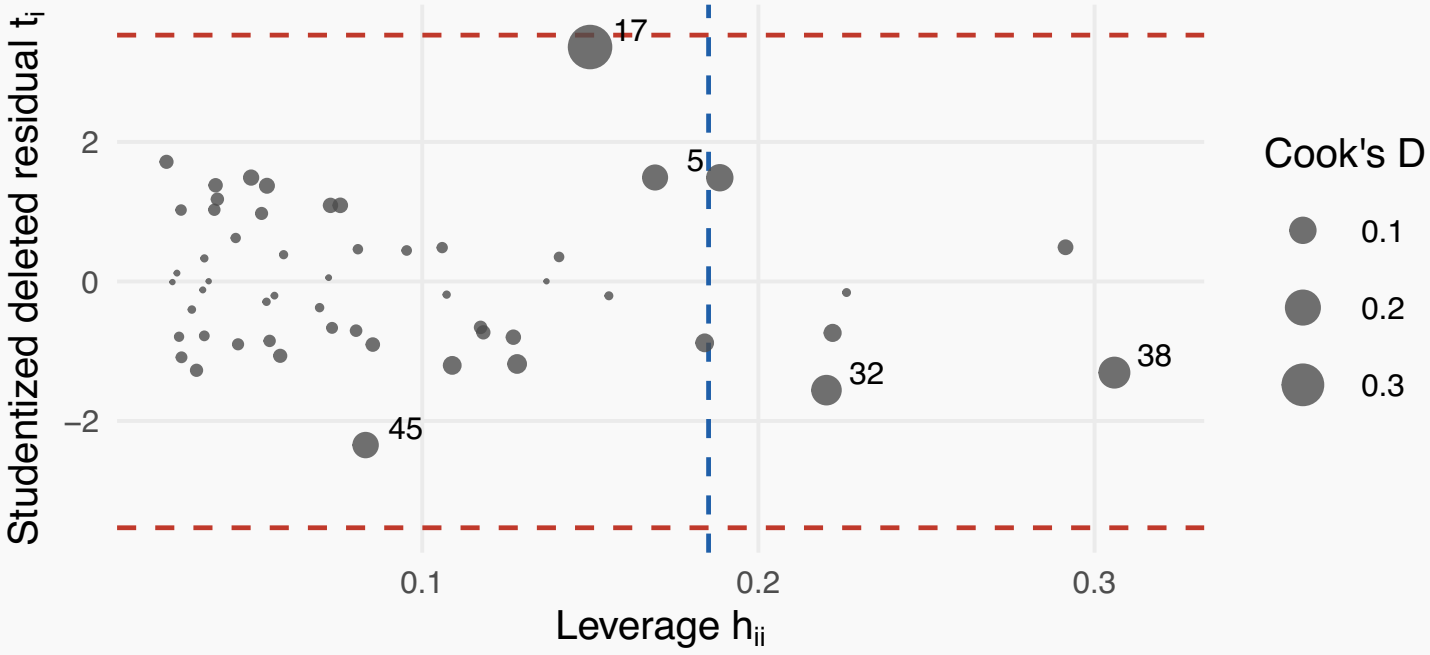
| Case | $e_i$ | $h_{ii}$ | $t_i$ | $DFFITS_i$ | $D_i$ |
|---|---|---|---|---|---|
| 1 | $0.069$ | 0.035 | $0.33$ | $0.06$ | 0.001 |
| 17 | $0.594$ | 0.150 | $3.36$ | $1.41$ | 0.329 |
| 38 | $-0.228$ | 0.306 | $-1.31$ | $-0.87$ | 0.148 |
| 45 | $-0.453$ | 0.083 | $-2.35$ | $-0.71$ | 0.091 |

# Exercise: solution

| Case | $e_i$ | $h_{ii}$ | $t_i$ | $DFFITS_i$ | $D_i$ |
|---|---|---|---|---|---|
| 1 | $0.069$ | 0.035 | $0.33$ | $0.06$ | 0.001 |
| 17 | $0.594$ | 0.150 | $3.36$ | $1.41$ | 0.329 |
| 38 | $-0.228$ | 0.306 | $-1.31$ | $-0.87$ | 0.148 |
| 45 | $-0.453$ | 0.083 | $-2.35$ | $-0.71$ | 0.091 |

- No outlier in $Y$: the largest $|t_i|$, 3.36 (case 17), is below 3.53.
- Case 38 has a high leverage ($0.31 > 0.185$) but a modest residual: not influential.
- Case 17: $DFFITS = 1.41 > 1$, it moves its own fitted value. But $D = 0.33$ is below the 20th percentile of $F(5, 49)$: little influence on the fit as a whole.

# All 54 patients at once




Red: the Bonferroni limits $\pm 3.53$. Blue: $2p/n$. The size of a point is its Cook's distance.

# In R

```
fit <- lm(log(Y) ~ X1 + X2 + X3 + X8, data = surg)
n <- nrow(surg); p <- length(coef(fit))
t_i <- rstudent(fit)
which(abs(t_i) > qt(1 - 0.05 / (2 * n), n - p - 1))
h <- hatvalues(fit); which(h > 2 * p / n)
which(abs(dffits(fit)) > 1)
D <- cooks.distance(fit); round(pf(D, p, n - p), 2)
which(abs(dfbetas(fit)) > 1, arr.ind = TRUE)
```

# In R

```
fit <- lm(log(Y) ~ X1 + X2 + X3 + X8, data = surg)
n <- nrow(surg); p <- length(coef(fit))
t_i <- rstudent(fit)
which(abs(t_i) > qt(1 - 0.05 / (2 * n), n - p - 1))
h <- hatvalues(fit); which(h > 2 * p / n)
which(abs(dffits(fit)) > 1)
D <- cooks.distance(fit); round(pf(D, p, n - p), 2)
which(abs(dfbetas(fit)) > 1, arr.ind = TRUE)
```

All of them come from the one fit. `pf(D, p, n - p)` gives the percentile of each $D_i$ in $F(p, n-p)$; `arr.ind = TRUE` gives the row (the case) and the column (the coefficient).

# What to do with an unusual case

- First check the data: a typing error, a measurement problem? Correct it if you can.
- Then ask whether the case belongs to the population studied (a different kind of patient, a faulty batch). If not, remove it, and say so.
- Otherwise keep it. Refit without it, and report whether the conclusions change.

# What to do with an unusual case

- First check the data: a typing error, a measurement problem? Correct it if you can.
- Then ask whether the case belongs to the population studied (a different kind of patient, a faulty batch). If not, remove it, and say so.
- Otherwise keep it. Refit without it, and report whether the conclusions change.

Never delete a case only because it does not fit: the model must describe the data, not the other way round.

# What to do with an unusual case

- First check the data: a typing error, a measurement problem? Correct it if you can.
- Then ask whether the case belongs to the population studied (a different kind of patient, a faulty batch). If not, remove it, and say so.
- Otherwise keep it. Refit without it, and report whether the conclusions change.

Never delete a case only because it does not fit: the model must describe the data, not the other way round.

The surgical unit: without case 17, every coefficient keeps its sign and stays significant. We keep the case, and mention it.

# Multicollinearity

# Informal diagnostics for multicollinearity

Signs that the predictors are strongly related to each other:

1. The estimated coefficients change a lot when a predictor is added or removed.
2. Predictors known to matter have nonsignificant $t$ tests.
3. A coefficient has the opposite sign of what theory or experience leads us to expect.
4. The standard errors of the coefficients are large.
5. $R^2$ changes little when a predictor is added or removed.
6. The predictors are highly correlated with each other.

# Informal diagnostics for multicollinearity

Signs that the predictors are strongly related to each other:

1. The estimated coefficients change a lot when a predictor is added or removed.
2. Predictors known to matter have nonsignificant $t$ tests.
3. A coefficient has the opposite sign of what theory or experience leads us to expect.
4. The standard errors of the coefficients are large.
5. $R^2$ changes little when a predictor is added or removed.
6. The predictors are highly correlated with each other.

Already seen: the raw power cells (Lecture 18: significant as a whole, no coefficient alone), and the cement data ($r(x_2, x_4) = -0.97$). The next slide gives a number.

# The variance inflation factor

Regress $X_k$ on the other predictors, with coefficient of determination $R_k^2$:

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad Var(\hat{\beta}_k) = VIF_k \times \frac{\sigma^2}{\sum_i (X_{ik} - \bar{X}_k)^2}.$$

$VIF_k$ is the inflation of $Var(\hat{\beta}_k)$ due to the other predictors (Lecture 16: $1/(1 - r^2)$ with two predictors).

# The variance inflation factor

Regress $X_k$ on the other predictors, with coefficient of determination $R_k^2$:

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad Var(\hat{\beta}_k) = VIF_k \times \frac{\sigma^2}{\sum_i (X_{ik} - \bar{X}_k)^2}.$$

$VIF_k$ is the inflation of $Var(\hat{\beta}_k)$ due to the other predictors (Lecture 16: $1/(1 - r^2)$ with two predictors).

- Serious multicollinearity: the largest $VIF$ above 10, or a mean $VIF$ well above 1.
- The cement data: 38.5, 254.4, 46.9, 282.5. This is why the three searches disagreed.
- The surgical unit ($X_1, X_2, X_3, X_8$): 1.10, 1.02, 1.05, 1.09. No problem.

# The variance inflation factor

Regress $X_k$ on the other predictors, with coefficient of determination $R_k^2$:

$$VIF_k = \frac{1}{1 - R_k^2}, \qquad Var(\hat{\beta}_k) = VIF_k \times \frac{\sigma^2}{\sum_i (X_{ik} - \bar{X}_k)^2}.$$

$VIF_k$ is the inflation of $Var(\hat{\beta}_k)$ due to the other predictors (Lecture 16: $1/(1 - r^2)$ with two predictors).

- Serious multicollinearity: the largest $VIF$ above 10, or a mean $VIF$ well above 1.
- The cement data: 38.5, 254.4, 46.9, 282.5. This is why the three searches disagreed.
- The surgical unit ($X_1, X_2, X_3, X_8$): 1.10, 1.02, 1.05, 1.09. No problem.

In R: `car::vif(fit)`, from the package `car`.

# Summary

- Search procedures move one predictor at a time. Stepwise regression can remove what it added; with correlated predictors, the procedures can disagree.
- Validation: predict cases used neither to choose nor to fit the model ($MSPE$).
- Outlier in $Y$: studentized deleted residual $t_i$, with Bonferroni. Outlier in $X$: leverage $h_{ii} > 2p/n$.
- Influence: $DFFITS$ (its own fit), Cook's $D$ (all the fit), $DFBETAS$ (each coefficient). Check, understand, refit without, report.
- $VIF_k = 1/(1 - R_k^2)$: above 10, serious multicollinearity.