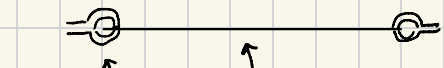
10/20/2025 Categorical predictors (part 2)

Example 1 (Strength of worsted yarn)



1 cycle { 1) Pull until yarn extends by, e.g. 10 mm ("amplitude")
2) Relax back to original length

⇝ Repeat ① + ② until yarn breaks & record # cycles

Data:

```
> head(Wool)
  len amp load cycles
1 250   8   40    674
2 250   8   45    370
3 250   8   50    292
4 250   9   40    338
5 250   9   45    266
6 250   9   50    210
```

good design

Variables

len = length of yarn (250, 300, 350 mm)

amp = Amplitude of cycle (8, 9, 10 mm)

load = load put on yarn (40, 45, 50 g)

cycles = # cycles until yarn fails

Treat each predictor as a categorical variable, a.k.a "factor variable" and let $y = \log(\text{cycles})$ (response)

"Main effects" model: $y = \beta_0 + \beta_1 \text{len}_{300} + \beta_2 \text{len}_{350} + \beta_3 \text{amp}_9 + \beta_4 \text{amp}_{10} + \beta_5 \text{load}_{45} + \beta_6 \text{load}_{50} + \varepsilon$

(7 parameters)

```
> lmod1 <- lm(log(cycles)~len+amp+load,wool.df)
> X <- model.matrix(lmod1)
> head(X)
  (Intercept) len300 len350 amp9 amp10 load45 load50
1           1      0      0    0     0      0      0
2           1      0      0    0     0      1      0
3           1      0      0    0     0      0      1
4           1      0      0    1     0      0      0
5           1      0      0    1     0      1      0
6           1      0      0    1     0      0      1
```

e.g. $\beta_1 = E[y \mid \text{len}=300, \text{amp}, \text{load}] - E[y \mid \text{len}=250, \text{amp}, \text{load}]$

(Same for amp = 8, 9, 10 load = 40, 45, 50)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.48287    0.09644  67.225  < 2e-16 ***
len300       0.91833    0.08928  10.286 1.97e-09 ***
len350       1.66477    0.08928  18.646 4.10e-14 ***
amp9        -0.65521    0.08928  -7.339 4.31e-07 ***
amp10       -1.26173    0.08928 -14.132 7.19e-12 ***
load45      -0.32529    0.08928  -3.643  0.00162 **
load50      -0.78524    0.08928  -8.795 2.62e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1894 on 20 degrees of freedom
Multiple R-squared:  0.9691,    Adjusted R-squared:  0.9598
F-statistic: 104.5 on 6 and 20 DF,  p-value: 4.979e-14
```

$$Var(\hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{jj}$$

$(X^T X)^{-1}$ is nearly constant along its diagonal

```
> round(diag(solve(t(X) %*% X)),3)
(Intercept)      len300      len350        amp9       amp10      load45      load50
      0.259       0.222       0.222       0.222       0.222       0.222       0.222
```

"Second order" model (Two-way interactions)

$$y = \beta_0 + \beta_1 len_{300} + \beta_2 len_{350} + \beta_3 amp_9 + \beta_4 amp_{10} + \beta_5 load_{45} + \beta_6 load_{50}$$
$$+ \beta_7 len_{300} amp_9 + \beta_8 len_{300} amp_{10} + \cdots + \beta_? amp_{10} load_{50} + \varepsilon$$

Q: How many parameters? $7 + 3 \times 4 = 19$

```
> tab1 <- tapply(wool.df$log_cycles, list(wool.df$amp, wool.df$load), mean)
> tab2 <- tapply(wool.df$log_cycles, list(wool.df$load, wool.df$len), mean)
> tab3 <- tapply(wool.df$log_cycles, list(wool.df$len, wool.df$amp), mean)
> round(tab1,2); round(tab2,2); round(tab3,2)
```

amp × load

amp \ load	40	45	50
8	7.32	7.02	6.58
9	6.70	6.33	5.92
10	6.09	5.79	5.26

load × length

load \ length	250	300	350
40	5.82	6.76	7.53
45	5.42	6.44	7.28
50	5.17	5.98	6.61

length × amp

length \ amp	8	9	10
250	6.03	5.58	4.80
300	6.93	6.48	5.76
350	7.96	6.89	6.57

e.g. 6.09 = Average log(cycle) among measurements with amp = 10, load = 40

e.g. Interaction len * amp means: the effect of length on yarn strength depends on amplitude

More specific:

$\beta_1 + \beta_7 = E[y \mid len = 300, amp = 9, load] - E[y \mid len = 250, amp = 9, load]$

$\beta_1 = E[y \mid len = 300, amp = 8, load] - E[y \mid len = 250, amp = 8, load]$

$\Rightarrow \beta_7$ = difference in length contrasts when amp = 9 vs amp = 8

Table 1 suggests no interaction between amp & load

Table 3 suggests possible amp * length interaction

```
> lmod1 <- lm(log(cycles)~len+amp+load,wool.df)
> lmod2 <- lm(log(cycles)~len+amp*load,wool.df)
> lmod3 <- lm(log(cycles)~len*amp+load,wool.df)
```

```
> anova(lmod1,lmod2)
Analysis of Variance Table

Model 1: log(cycles) ~ len + amp + load
Model 2: log(cycles) ~ len + amp * load
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     20 0.71742
2     16 0.70283  4  0.014597 0.0831 0.9864
```

```
> anova(lmod1,lmod3)
Analysis of Variance Table

Model 1: log(cycles) ~ len + amp + load
Model 2: log(cycles) ~ len * amp + load
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     20 0.71742
2     16 0.31626  4   0.40116 5.0737 0.007806 **
```

Example 2: Collaborative Learning (CL)

4 sections of Stat220, 2 instructors (Dan + Fei)

2 with CL, 2 without CL

CL, no-CL, CL, no-CL

- CL ⇔ mandatory attendance to a weekly 80 minute problem session (outside of class)
- Same HW, slides, exams
- Pre-course survey / quiz (assess pre-existing knowledge + social network)

Data: n = 258 students

```
> CL.df <- read.csv("CLdata.csv")
> head(CL.df)
  section    CL social midterm final_exam gpa
1       3  TRUE   TRUE   0.565      0.845   3
2       3  TRUE   TRUE   0.710      0.755   4
3       2 FALSE   TRUE   1.000      0.900   4
4       4 FALSE   TRUE   0.920      0.990   4
5       1  TRUE   TRUE   0.990      1.020   4
6       4 FALSE   TRUE   1.000      0.940   3
```

Each row has data for one student

- $Social_i$ = True if student i knows at least one other student enrolled in Stat 220, False otherwise
- gpa_i = gpa of student i (rounded to nearest integer) at start of quarter
- CL_i = True if student i enrolled in a CL section of Stat 220, False otherwise

Q: Does attendance to CL workshops improve student performance? Who benefits from CL?

Two-sample comparison	CL	no-CL
Average final score	83.7%	83.8%
Average midterm score	78.6%	79.1%
Avg (rounded) GPA	3.68	3.73
Proportion of social = True	79.0%	84.2%

~ should control for academic ability & social network

Consider: final ~ mt + gpa + CL * social

(mt, gpa: quantitative; CL, social: categorical)

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.27818    0.06082   4.574 7.53e-06 ***
midterm          0.36749    0.04919   7.470 1.32e-12 ***
gpa              0.06832    0.01827   3.740 0.000228 ***
socialTRUE       0.02465    0.02997   0.823 0.411530
CLTRUE           0.03753    0.03708   1.012 0.312463
socialTRUE:CLTRUE -0.04808   0.04098  -1.173 0.241848
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1251 on 251 degrees of freedom
Multiple R-squared:  0.3635,    Adjusted R-squared:  0.3508
F-statistic: 28.67 on 5 and 251 DF,  p-value: < 2.2e-16
```

Q: Which students benefit most from CL?

CL \ social	no social	social
no CL	0	2.47 %
CL	3.75 %	2.47 + 3.75 - 4.81 = 1.4 %

Interpretation: Among students knowing 0 others at the class at start of quarter, those in CL sections scored 3 to 4 percent higher on the final exam than those in non-CL sections (but CI contains 0)

~ Weak evidence of CL having an effect

Repeat the experiment? (Winter 2026!)

10/22 Model Diagnostics (Part 1)

Motivating example: 4 artificial datasets

```
anscombe.df <- read.csv("anscombe.csv")
anscombe.df
```

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

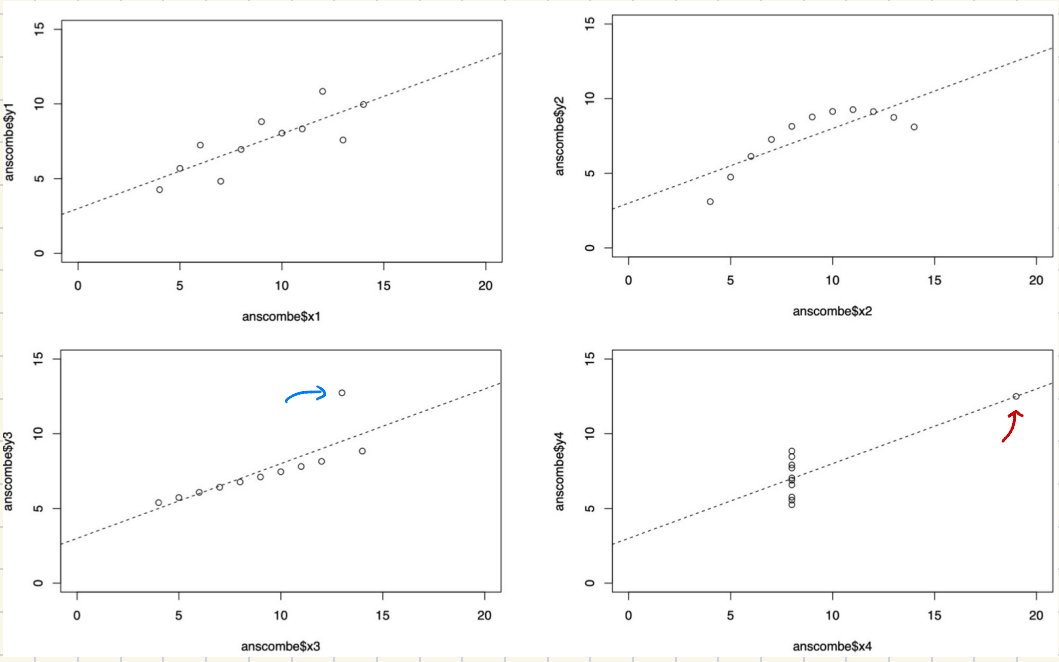
Fit each separately $y_i \sim x_i$ $i \le 4$

$\rightsquigarrow$ identical fitted models

$\hat{\beta}_0 = 3$, $\hat{\beta}_1 = 1/2$

$\hat{\sigma}^2 = 13.75$, $R^2 = 0.667$

Anscombe, 1973



Outlier: A case far from the fitted regression line

Influential point: A case that greatly influences the model fit

Goal: Diagnose whether the model fit was appropriate

Our main tool is the residual vector $e = y - \hat{y}$

Residuals

Recall: matrix formulation $y = X\beta + \varepsilon$ $\quad \text{Var}(\varepsilon) = \sigma^2 I_n$

where X is full rank (no column is a linear combination of others)

Hat matrix: $H = X(X^TX)^{-1}X^T$ $\quad (n \times n$ matrix$)$

↳ Claim: $Hy = \hat{y}$ "y-hat"

Proof: $Hy = X\underbrace{(X^TX)^{-1}X^Ty}_{\hat{\beta}} = X\hat{\beta} = \hat{y}$ (fitted values)

$\hat{\beta}$ (OLS estimator) □

$e = y - \hat{y} = (I - H)y$

Result: If $\varepsilon_1, \ldots, \varepsilon_n$ are independent and $\text{Var}(\varepsilon_i) = \sigma^2$, then residuals $e_1, \ldots, e_n$ satisfy

$$E[e_i] = 0 \quad \text{and} \quad \text{Var}(e_i) = \sigma^2(1 - h_{ii})$$

↳ i^{th} diagonal entry of H

Residuals vs Errors

Same: mean zero, normally distributed if $\varepsilon \sim N(0, \sigma^2 I_n)$

different: residuals are observed, errors aren't.

$\text{Var}(e_i)$ depends on i, $\text{Var}(\varepsilon_i) = \sigma^2$ for all i

$\{e_i\}$ are correlated ($\sum_{i=1}^{n} e_i = 0$), $\{\varepsilon_i\}$ are independent

Leverage: h_{ii} is called the leverage of case i

Special case: When $p=1$, $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$ (SLR)

↑ HW problem

As x_i moves away from the average $\bar{x}$, it has higher leverage, e.g. the point in Anscombe's 4th dataset determines $\hat{\beta}_1$ (without it, $\hat{\beta}_1$ is undefined)

General Properties

- $\frac{1}{n} \leq h_{ii} \leq 1$
- $\sum_{i=1}^{n} h_{ii} = p+1$ where p = # predictors
- Each row and column of H sums to 1 ★

$$\left(\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1 \right)$$

$$\hat{y} = Hy \Rightarrow \hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

High leverage: As $h_{ii} \to 1$, $\hat{y}_i \to y_i$ because of ★

Outliers: Atypical cases that don't seem to follow the same model as rest of data, i.e. an observation with large residual

Suppose case i is a suspected outlier $\rightsquigarrow$ Formally, want to test: $H_{0i}: E[y_i] = x_i^T \beta$

Q: Test statistic: $(y_i - x_i^T \hat{\beta}) / se(y_i - x_i^T \hat{\beta})$?

If case i is an outlier, then $x_i^T \hat{\beta}$ could be a poor estimate of $x_i^T \beta$. Instead, consider:

Outlier test procedure ★

1) Delete case i from the data

2) Estimate β and σ^2 $\rightsquigarrow$ $\hat{\beta}_{(-i)}$ and $\hat{\sigma}^2_{(-i)}$

3) Compute $\hat{y}_{i(-i)} = x_i^T \hat{\beta}_{(-i)}$ (fitted value for deleted case)

Since y_i is independent of other y_j's $j \neq i$, we have

$$Var(y_i - \hat{y}_{i(-i)}) = \sigma^2 + \sigma^2 x_i^T (X_{(-i)}^T X_{(-i)})^{-1} x_i$$

$$\approx \hat{\sigma}^2_{(-i)} (1 + x_i^T (X_{(-i)}^T X_{(-i)})^{-1} x_i)$$

4) T-statistic: $t_i = \dfrac{y_i - \hat{y}_{i(-i)}}{\hat{\sigma}_{(-i)} \sqrt{1 + x_i^T (X_{(-i)}^T X_{(-i)}) x_i}}$ $\sim$ Student-t

$df = n - 1 - (p+1)$

Result: $t_i = \dfrac{e_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}} = r_i \left(\dfrac{n-p-2}{n-(p+1)-r_i^2} \right)^{1/2}$ where

$\overset{H_0}{\sim}$ Student-t, $df = n-p-2$

$r_i = e_i / (\hat{\sigma} \sqrt{1-h_{ii}})$

Details for this result can be found in Appendix A.13 in ALR textbook

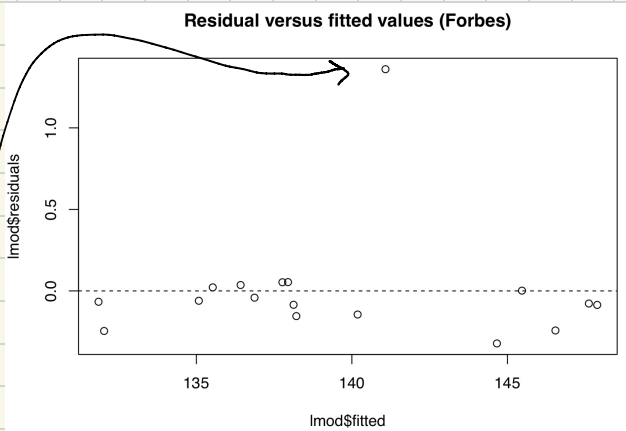
Choosing the significance level

e.g. Forbes data

Residuals e vs $\hat{y}$

log(pressure) ~ bp

Case 12 is the suspected outlier



Q: Is it valid to perform the test ★ at level, e.g. $\alpha = 5\%$, for case 12? Slight complication...

- Before seeing the data, we would not have known case 12 to be the suspected outlier (all 17 cases on equal footing)
- $|t_{12}|$ happens to be most extreme among $|t_1|, \ldots, |t_{17}|$
- Our actual test statistic: $\max_{i=1,\ldots,17} |t_i|$ (which happens to be $|t_{12}|$)

Bonferroni correction

Suppose $n = 65$, $p = 3$ ⇝ qt(0.975, 65 - 3 - 2) = 2

For each i, the outlier test has $P_0(|t_i| > 2) = 0.05$

However, $P_0(\max_{i=1,\dots,65} |t_i| > 2) \approx 96\% >> 5\%$ for independent (t_i)

↑ meaning P if H_0 is true

"Union" or "Bonferroni" bound: $P_0(\max_{i \le n} |t_i| > 2) \le n \underbrace{P_0(|t_i| > 2)}_{5\%}$

⇒ To control type I error at 5%, replace 2 with the t-quantile when $\alpha' = 0.05/n$ ($= 0.00077$ if $n = 65$)

↳ $qt(1 - \alpha'/2, 60) = 3.54$

$\xRightarrow{\text{union bd}} P_0(\max_{i \le 65} |t_i| > 3.54) \le 0.05$

e.g. Forbes

```
lmod <- lm(lpres~bp,df)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.13778    3.34020  -12.62 2.18e-09 ***
bp            0.89549    0.01645   54.43  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.379 on 15 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.9946
F-statistic:  2963 on 1 and 15 DF,  p-value: < 2.2e-16
```

$r_i = \frac{e_i}{\hat\sigma\sqrt{1-h_{ii}}} = \frac{1.36}{0.379\sqrt{1-0.064}} = 3.71$

$t_i = 3.71\left(\frac{17-1-2}{17-2-3.71^2}\right)^{1/2} \approx 12.4$

```
> round(residuals(lmod),3)
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
-0.246 -0.067 -0.060  0.022  0.036 -0.041  0.053  0.054 -0.155 -0.085 -0.145  1.360  0.002 -0.322 -0.243
   16    17
-0.077 -0.086
> round(hatvalues(lmod),3)
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17
0.193 0.200 0.107 0.098 0.083 0.076 0.067 0.065 0.063 0.064 0.060 0.064 0.140 0.119 0.172 0.210 0.220
```

R shortcut:

```
> round(rstudent(lmod),3)
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15
-0.710 -0.190 -0.163  0.059  0.097 -0.110  0.140  0.142 -0.410 -0.225 -0.382 12.407  0.005 -0.900 -0.691
   16    17
-0.222 -0.249
```

$qt(1 - 0.05/(2 \cdot 17), df = 17-3) = 3.59 < t_{12}$ ⇝ case 12 is an outlier

Equivalent: `17 * 2 * pt(12.41, df = 14, lower = FALSE) < 0.05`

$\rightsquigarrow$ Refit without case 12

```
> lmod2 <- lm(lpres~bp,forbes.df[-12,])
> summary(lmod2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.30838    1.00052  -41.29    5e-16 ***
bp            0.89099    0.00493  180.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1133 on 14 degrees of freedom
Multiple R-squared:  0.9996,    Adjusted R-squared:  0.9995
F-statistic: 3.266e+04 on 1 and 14 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_1$ went down $\sim$ 1 se

R^2 went up

Q: How do we quantify the influence of a case (eg. case 12 in Forbes data) on the model fit (e.g. $\hat{Y}$ or $\hat{\beta}$)?

For case i, define it's Cook's distance D_i

$$D_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})^T (X^T X)(\hat{\beta}_{(-i)} - \hat{\beta})}{(p+1)\hat{\sigma}^2} = \frac{1}{p+1} r_i^2 \frac{h_{ii}}{1-h_{ii}}$$

↑ see App. A.13 for details

$$= \frac{\| \hat{y} - \hat{y}_{(-i)} \|^2}{(p+1)\hat{\sigma}^2}$$

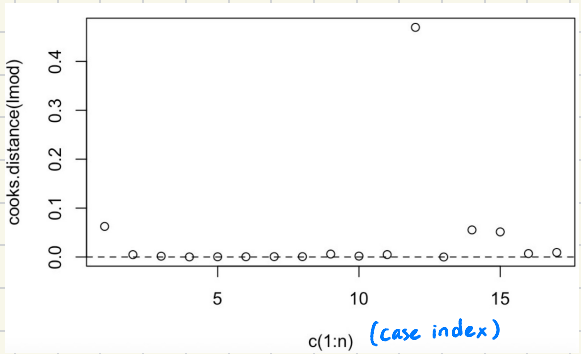
Intuition: D_i is large when r_i^2 and h_{ii} are large

↑ Standardized residual^2

↑ leverage (how far x_i is from the bulk)

In R:

```
> n <- nrow(forbes.df)
> plot(c(1:n),cooks.distance(lmod))
> abline(h=0,lty=2)
```



Plot of Cook's D versus case #

"Large Cook's D" ≥ 1 (rule of thumb)

↖ "Index plot"

Example (Rat data)

```
> rat.df <- read.csv("rat.csv")
> head(rat.df)
  BodyWt LiverWt Dose    y
1    176     6.5 0.88 0.42
2    176     9.5 0.88 0.25
3    190     9.0 1.00 0.56
4    176     8.9 0.88 0.23
5    200     7.2 1.00 0.23
6    167     8.9 0.83 0.32
```

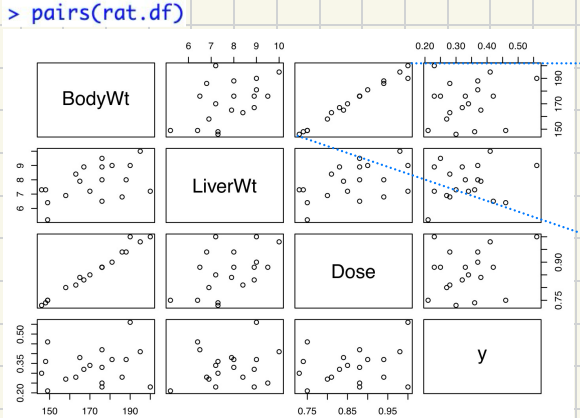
n = 19 rats, each is administered a drug with Dose ≈ proportional to body weight

(large livers absorb more than small livers)

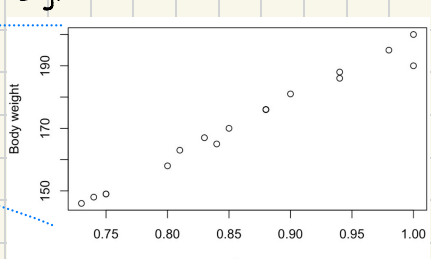
After some time ⇝ murine sacrifice + body measurements are taken. The response y = % of dose in liver at death

Null hypothesis: After the fixed amount of time, y has no relationship with X = (body weight, liver weight, dose)

```
> pairs(rat.df)
```



e.g.



⇝ y appears $\approx$ unrelated to each variable individually

but...

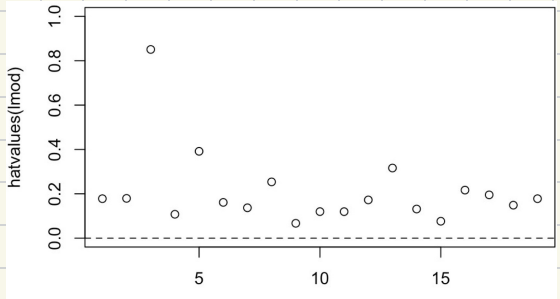
```
> lmod <- lm(y~BodyWt+LiverWt+Dose,rat.df)
> summary(lmod)
```

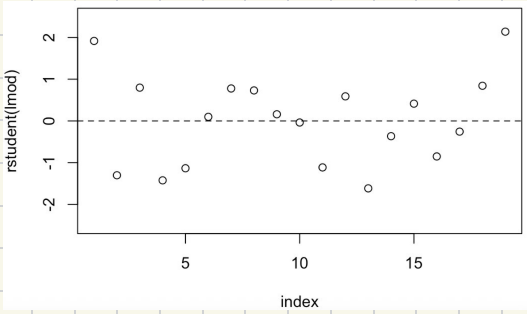
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.265922   0.194585   1.367   0.1919
BodyWt      -0.021246   0.007974  -2.664   0.0177 *
LiverWt      0.014298   0.017217   0.830   0.4193
Dose         4.178111   1.522625   2.744   0.0151 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07729 on 15 degrees of freedom
Multiple R-squared:  0.3639,    Adjusted R-squared:  0.2367
F-statistic:  2.86 on 3 and 15 DF,  p-value: 0.07197
```

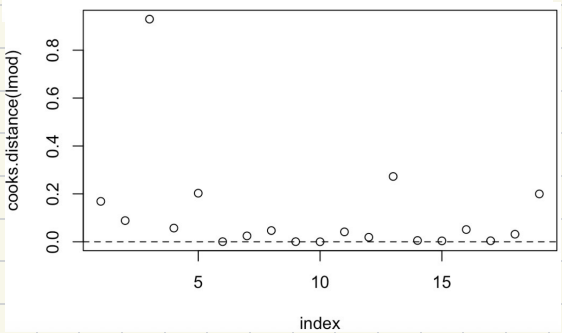
Interpretation: Neither body weight nor dose has relationship with y ignoring the other, but together, they have an association with y

Paradox: Dose is determined by body weight, so how can they be more informative in combination for predicting y?

```
index <- c(1:nrow(rat.df))
plot(index,hatvalues(lmod),ylim=c(0,1))
abline(h=0,lty=2)
plot(index,rstudent(lmod),ylim=c(-2.5,2.5))
abline(h=0,lty=2)
plot(index,cooks.distance(lmod))
abline(h=0,lty=2)
```

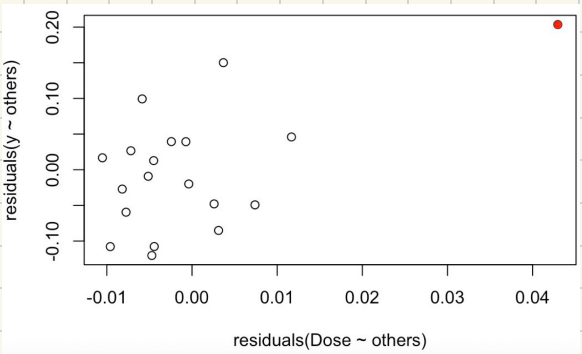






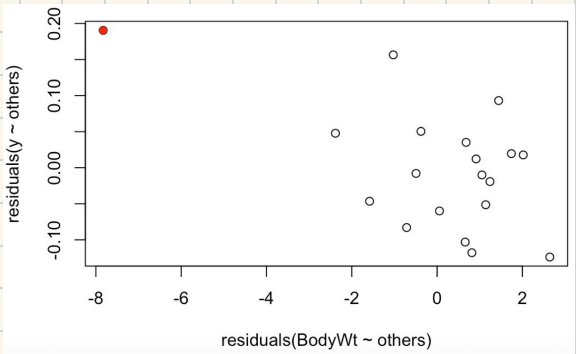
Diagnostic plots suggest case 3 is influential

```
> lmod3 <- lm(y~BodyWt+LiverWt,rat.df)
> tmp3 <- lm(Dose~BodyWt+LiverWt,rat.df)
> plot(residuals(tmp3),residuals(lmod3),
+      xlab="residuals(Dose ~ others)",
+      ylab="residuals(y ~ others)")
> points(residuals(tmp3)[3],residuals(lmod3)[3],col="red",pch=16)
```



body & liver weight

```
> lmod4 <- lm(y~LiverWt+Dose,rat.df)
> tmp4 <- lm(BodyWt~Dose+LiverWt,rat.df)
> plot(residuals(tmp4),residuals(lmod4),
+      xlab="residuals(BodyWt ~ others)",
+      ylab="residuals(y ~ others)")
> points(residuals(tmp4)[3],residuals(lmod4)[3],col="red",pch=16)
```



Dose & liver weight

Refit model excluding case 3:

```
> lmod2 <- lm(y~BodyWt+LiverWt+Dose,rat.df[-3,])
> summary(lmod2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.311427   0.205094   1.518    0.151
BodyWt      -0.007783   0.018717  -0.416    0.684
LiverWt      0.008989   0.018659   0.482    0.637
Dose         1.484877   3.713064   0.400    0.695

Residual standard error: 0.07825 on 14 degrees of freedom
Multiple R-squared:  0.02106,   Adjusted R-squared:  -0.1887
F-statistic: 0.1004 on 3 and 14 DF,  p-value: 0.9585
```

} nothing is significant

~> The paradox is resolved (spurious relation between y and (dose, bodyweight) is due to case 3)

Q: Toss case 3? Look into details of this case to decide. (Perhaps dose administration did not follow protocol)