Fitting Linear Models

DAAG Chapter 6
interpretation of coefficients can be tricky…

```
> litters
  lsize bodywt brainwt
1   3  9.447   0.444
2   3  9.780   0.436
3   4  9.155   0.417
4   4  9.613   0.429
5   5  8.850   0.425
6   5  9.610   0.434
7   6  8.298   0.404
8   6  8.543   0.439
9   7  7.400   0.409
10  7  8.335   0.429
11  8  7.040   0.414
12  8  7.253   0.409
13  9  6.600   0.387
14  9  7.260   0.433
15 10  6.305   0.410
16 10  6.655   0.405
17 11  7.183   0.435
18 11  6.133   0.407
19 12  5.450   0.368
20 12  6.050   0.401
> pairs(litters)
```
titl <- "Brain Weight as a Function of Litter Size and Body Weight"
plot(litters[, 1], litters[, 3], xlab = "Litter Size", ylab = "Brain Weight (g)", pch = 16)
u1 <- lm(brainwt ~ lsize, data = litters)
abline(u1)
mtext(side = 3, line = 0.25, text = paste("brainwt =", round(u1$coef[1], 2), "+",
        round(u1$coef[2], 5), "[SE =", round(se, 5), "]", "lsize"), cex = 1.0)
plot(litters[, 2], litters[, 3], xlab = "Body Weight (g)", ylab = "Brain Weight (g)", pch = 16)
u2 <- lm(brainwt ~ bodywt, data = litters)
abline(u2)
mtext(side = 3, line = 0.25, text = paste("brainwt =", round(u2$coef[1], 2), "+",
        round(u2$coef[2], 4), "bodywt"), cex = 1.0)
plot(litters[, 1], litters[, 2], xlab = "Litter Size", ylab = "Body Weight (g)", pch = 16, mkh = 0.04)
r3 <- cor(litters[, 1], litters[, 2])
mtext(side = 3, line = 0.25, text = paste("Correlation =", round(r3, 2)
        ), cex = 1.0)
u <- lm(brainwt ~ lsize + bodywt, data = litters)
hat <- fitted(u)
plot(hat, litters[, 3], xlab = "Fitted Weight (g)", ylab = "Brain Weight (g)", pch = 1, lwd=2)
se <- summary(u)$coef[2, 2]
se1 <- summary(u)$coef[3, 2]
mtext(side = 3, line = 0.5,
        text = paste("brainwt =",
        round(u$ coef[1], 2), "+", round(u$coef[2], 4),
        "[SE =", round(se, 4), "] ","lsize \n+",
        round(u$coef[3], 4), "[SE =", round(se1, 4), "] ",
        "bodywt",sep=""), cex = 1.0)
> summary(u)

Call:
lm(formula = brainwt ~ lsize + bodywt, data = litters)

Residuals:
   Min      1Q  Median      3Q     Max
-0.0230005 -0.0098821  0.0004512  0.0092036  0.0180760

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)  0.1782472   0.0753233    2.366  0.03010 *
lsize       0.0066900   0.0031322    2.136  0.04751 *
bodywt      0.0243064   0.0067786    3.586  0.00228 **

---

Residual standard error: 0.01195 on 17 degrees of freedom
Multiple R-Squared: 0.6505,  Adjusted R-squared: 0.6094
F-statistic: 15.82 on 2 and 17 DF,  p-value: 0.0001315

estimated amount by which E[Y] increases as lsize goes up by one, holding bodywt constant
"Simpson's Paradox"

- 2 X 2 table analysis ignores effects of drug-drug association on drug-AE association

<table>
<thead>
<tr>
<th></th>
<th>Rosinex</th>
<th>No Rosinex</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nausea</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ganclex</td>
<td>81</td>
<td>9</td>
<td>82</td>
</tr>
<tr>
<td>No Nausea</td>
<td>1</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>82</td>
<td>10</td>
<td>92</td>
</tr>
<tr>
<td>RR</td>
<td>1</td>
<td>1</td>
<td>4.58</td>
</tr>
<tr>
<td>Nausea</td>
<td>810</td>
<td>99</td>
<td>909</td>
</tr>
</tbody>
</table>
berkeley admission's data 1973

<table>
<thead>
<tr>
<th></th>
<th>Applicants</th>
<th>% admitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>8442</td>
<td>44%</td>
</tr>
<tr>
<td>Women</td>
<td>4321</td>
<td>35%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Major</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Applicants</td>
<td>% admitted</td>
</tr>
<tr>
<td>A</td>
<td>825</td>
<td>62%</td>
</tr>
<tr>
<td>B</td>
<td>560</td>
<td>63%</td>
</tr>
<tr>
<td>C</td>
<td>325</td>
<td>37%</td>
</tr>
<tr>
<td>D</td>
<td>417</td>
<td>33%</td>
</tr>
<tr>
<td>E</td>
<td>191</td>
<td>28%</td>
</tr>
<tr>
<td>F</td>
<td>272</td>
<td>6%</td>
</tr>
</tbody>
</table>
Bad Things Can Happen…

DATA

happiness

hours per week on studies
HAP = β₀ + β₁ \times \text{HOURS}, \beta₁ will be estimated to be negative
A 2\textsuperscript{nd} Look at the DATA

happiness

hours per week on studies
A 2\textsuperscript{nd} Look at the DATA

hours per week on studies

happiness
Other Odd Things Can Happen…
Other Odd Things Can Happen…
important diversion: these issues are related to "confounding"

but...

are not the same...
Confounding and Causality

• Confounding is a causal concept

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Population D</th>
<th>Population d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug (factual)</td>
<td>Not drug (counterfactual)</td>
</tr>
<tr>
<td>Y=1</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Y=0</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>a=0.3</td>
<td>b=0.2</td>
<td></td>
</tr>
</tbody>
</table>

True causal effect = \( \frac{a}{b} = 1.5 \) or \( \frac{a}{(1-a)} \div \frac{b}{(1-b)} = 1.71 \)
Estimated causal effect = \( \frac{a}{c} = 3 \) or \( \frac{a}{(1-a)} \div \frac{c}{(1-c)} = 3.86 \)

• “The association in the combined D+d populations is confounded for the effect in population D”
Why does this happen?

• For confounding to occur there must be some characteristics/covariates/conditions that distinguish D from d.

• However, the existence of such factors does not in and of itself imply confounding.

• For example, D could be males and d females but it could still be the case that \( b = c \).
# Stratification can introduce confounding

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Population D</th>
<th>Population d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug (actual)</td>
<td>Not drug (counter)</td>
</tr>
<tr>
<td>Y=1</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>Y=0</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>a=0.3</td>
<td>b=0.2</td>
<td>c=0.2</td>
</tr>
</tbody>
</table>

True causal effect = a-b = 0.1  
Estimated causal effect = a-c = 0.1  
No confounding

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Population D</th>
<th>Population d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug (actual)</td>
<td>Not drug (counter)</td>
</tr>
<tr>
<td>Y=1</td>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>Y=0</td>
<td>35</td>
<td>8</td>
</tr>
<tr>
<td>a=0.3</td>
<td>b=0.2</td>
<td>c=0.25</td>
</tr>
</tbody>
</table>

True = a-b = 0.1  
Estimated = a-c = 0.05  
Confounding

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Population D</th>
<th>Population d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Drug (actual)</td>
<td>Not drug (counter)</td>
</tr>
<tr>
<td>Y=1</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Y=0</td>
<td>35</td>
<td>72</td>
</tr>
<tr>
<td>a=0.3</td>
<td>b=0.2</td>
<td>0.1875</td>
</tr>
</tbody>
</table>

True = a-b =0.1  
Estimated = a-c = 0.1125  
Confounding
Non-Collapsibility without Confounding

<table>
<thead>
<tr>
<th>Population D</th>
<th>Drug (factual)</th>
<th>Not drug (counterfactual)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y=1</td>
<td>Y=0</td>
</tr>
<tr>
<td>Covariate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z=1</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>Z=0</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>80</td>
</tr>
</tbody>
</table>

True causal effect | Z=1: $0.8/0.2 \div 0.6/0.4 = 2.67$
True causal effect | Z=0: $0.4/0.6 \div 0.2/0.8 = 2.67$
True causal effect ignoring Z: $0.6/0.4 \div 0.4/0.6 = 2.25$

<table>
<thead>
<tr>
<th>Population d</th>
<th>Drug (factual)</th>
<th>Not drug (counterfactual)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y=1</td>
<td>Y=0</td>
</tr>
<tr>
<td>Covariate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z=1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z=0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>80</td>
</tr>
</tbody>
</table>

Estimated causal effect | Z=1: $0.8/0.2 \div 0.6/0.4 = 2.67$
Estimated causal effect | Z=0: $0.4/0.6 \div 0.2/0.8 = 2.67$
Estimated causal effect ignoring Z: $0.6/0.4 \div 0.4/0.6 = 2.25$
# Collapsibility with Confounding

## Population D

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Drug (factual)</th>
<th>Not drug (counterfactual)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y=1</td>
<td>Y=0</td>
</tr>
<tr>
<td>Z=1</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>Z=0</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>120</strong></td>
<td><strong>80</strong></td>
</tr>
</tbody>
</table>

True causal effect | $Z=1$: $0.8 / 0.2 \div 0.6 / 0.4 = 2.67$
True causal effect | $Z=0$: $0.4 / 0.6 \div 0.2 / 0.8 = 2.67$
True causal effect ignoring $Z$: $0.6 / 0.4 \div 0.4 / 0.6 = 2.25$

## Population d

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Drug (factual)</th>
<th>Not drug (counterfactual)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y=1</td>
<td>Y=0</td>
</tr>
<tr>
<td>Z=1</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Z=0</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>90</strong></td>
<td><strong>160</strong></td>
</tr>
</tbody>
</table>

Estimated causal effect | $Z=1$: $0.8 / 0.2 \div 0.6 / 0.4 = 2.67$
Estimated causal effect | $Z=0$: $0.4 / 0.6 \div 0.2 / 0.8 = 2.67$
Estimated causal effect ignoring $Z$: $0.6 / 0.4 \div 0.36 / 0.64 = 2.67$
### the hills example

```r
> hills

<table>
<thead>
<tr>
<th>Name</th>
<th>dist</th>
<th>climb</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenmantle</td>
<td>2.4</td>
<td>650</td>
<td>0.2680556</td>
</tr>
<tr>
<td>Carnethy</td>
<td>6.0</td>
<td>2500</td>
<td>0.8058333</td>
</tr>
<tr>
<td>Craig Dunain</td>
<td>6.0</td>
<td>900</td>
<td>0.5608333</td>
</tr>
<tr>
<td>Ben Rha</td>
<td>7.5</td>
<td>800</td>
<td>0.7600000</td>
</tr>
<tr>
<td>Ben Lomond</td>
<td>8.0</td>
<td>3070</td>
<td>1.0377778</td>
</tr>
<tr>
<td>Goatfell</td>
<td>8.0</td>
<td>2866</td>
<td>1.2202778</td>
</tr>
<tr>
<td>Bens of Jura</td>
<td>16.0</td>
<td>7500</td>
<td>3.4102778</td>
</tr>
<tr>
<td>Cairnpapple</td>
<td>6.0</td>
<td>800</td>
<td>0.6061111</td>
</tr>
<tr>
<td>Scolty</td>
<td>5.0</td>
<td>800</td>
<td>0.4958333</td>
</tr>
<tr>
<td>Traprain</td>
<td>6.0</td>
<td>650</td>
<td>0.6625000</td>
</tr>
<tr>
<td>Lairig Ghru</td>
<td>28.0</td>
<td>2100</td>
<td>3.2111111</td>
</tr>
<tr>
<td>Dollar</td>
<td>5.0</td>
<td>2000</td>
<td>0.7175000</td>
</tr>
<tr>
<td>Lomonds</td>
<td>9.5</td>
<td>2200</td>
<td>1.0833333</td>
</tr>
<tr>
<td>Cairn Table</td>
<td>6.0</td>
<td>500</td>
<td>0.7355556</td>
</tr>
<tr>
<td>Eildon Two</td>
<td>4.5</td>
<td>1500</td>
<td>0.4488889</td>
</tr>
<tr>
<td>Cairngorm</td>
<td>10.0</td>
<td>3000</td>
<td>1.2041667</td>
</tr>
<tr>
<td>Seven Hills</td>
<td>14.0</td>
<td>2200</td>
<td>1.6402778</td>
</tr>
<tr>
<td>Knock Hill</td>
<td>3.0</td>
<td>350</td>
<td>1.3108333</td>
</tr>
<tr>
<td>Black Hill</td>
<td>4.5</td>
<td>1000</td>
<td>0.2902778</td>
</tr>
<tr>
<td>Creag Beag</td>
<td>5.5</td>
<td>600</td>
<td>0.5427778</td>
</tr>
</tbody>
</table>
```
begin with scatterplot matrices...

```r
library(lattice)
splom(~hills, cex.labels=1.2,
    varnames=c("dist\n(miles)", "climb\n(feet)", "time\n(hours)"))

splom(~log(hills), cex.labels=1.2,
    varnames=c("dist\n(log miles)", "climb\n(log feet)",
    "time\n(log hours)"))
```

why log?
- perhaps expect prediction error to be a fraction of the predicted time (think: prediction of 15 minutes as against a prediction of 3 hours)
- long tails
- marathon record not 26 X 4 minutes...
plot(hills$dist, hills$time)
identify(hills$dist, hills$time)

> hills[18,]
     dist climb     time
Knock Hill 3 350 1.310833

actually an error - should be 0.31!
plot(lm(time ~ dist + climb, data = hills), which = 1:4)
plot(lm(time~dist+climb,data=hills,subset=-18),which=1:4)
how about an interaction?

logHills <- log(hills)
names(logHills) <- c("logDist", "logClimb", "logTime")

hillsInt.lm <- lm(logTime~logDist*logClimb, data=logHills,subset=-18)
summary(hillsInt.lm)
par(mfrow=c(2,2))
plot(hillsInt.lm,which=1:4)

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -2.47552 | 0.96528    | -2.565  | 0.0156 * |
| logDist        | 0.16854  | 0.49129    | 0.343   | 0.7339   |
| logClimb       | 0.06724  | 0.13540    | 0.497   | 0.6231   |
| logDist:logClimb | 0.09928 | 0.06530    | 1.520   | 0.1389   |
> summary(lm(logTime~logDist+logClimb, data=logHills, subset=-18))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -3.88155 | 0.28263    | -13.734 | 1.01e-14 *** |
| logDist        | 0.90924  | 0.06500    | 13.989  | 6.16e-15 *** |
| logClimb       | 0.26009  | 0.04839    | 5.375   | 7.33e-06 *** |
Stack Loss Example

> stackloss

<table>
<thead>
<tr>
<th>Air.Flow</th>
<th>Water.Temp</th>
<th>Acid.Conc.</th>
<th>stack.loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>27</td>
<td>89</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>27</td>
<td>88</td>
</tr>
<tr>
<td>3</td>
<td>75</td>
<td>25</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>62</td>
<td>24</td>
<td>87</td>
</tr>
<tr>
<td>5</td>
<td>62</td>
<td>22</td>
<td>87</td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>23</td>
<td>87</td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>24</td>
<td>93</td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>24</td>
<td>93</td>
</tr>
</tbody>
</table>

Oxidation of Ammonia to Nitric Acid on Successive Days

x1 airflow to the plant
x2 temperature of the cooling water
x3 concentration of nitric acid in the absorbing liquid

y $10 \times$ the percentage of ingoing ammonia that is lost as unabsorbed nitric acids
> m1 <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.)
> summary(m1)

Call:
  lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.)

Residuals:
   Min       1Q   Median       3Q      Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405

---

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-Squared: 0.9136, Adjusted R-squared: 0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
Some odd patterns and outlying points. How come?

1. Some points may be entirely erroneous
2. We might not have the best functional form
3. The random error might not be normal
4. Some points may reflect changing conditions - perhaps the plant requires time to reach equilibrium after significant input changes

focus on 1 and 2 to begin with. Lets try dropping #21.
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>50</td>
<td>19</td>
<td>72</td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>19</td>
<td>79</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>20</td>
<td>82</td>
</tr>
<tr>
<td>20</td>
<td>56</td>
<td>20</td>
<td>91</td>
</tr>
<tr>
<td>21</td>
<td>70</td>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>15</td>
</tr>
</tbody>
</table>

21 is the last day of measurement
big input change, no change in output…
```r
> m2 <- lm(stack.loss~Air.Flow+Water.Temp+Acid.Conc.,subset=-21)
> summary(m2)

Call:
  lm(formula = stack.loss ~ Air.Flow + Water.Temp + Acid.Conc.,
     subset = -21)

Residuals:
     Min      1Q  Median      3Q     Max
-3.0449 -2.0578  0.1025  1.0709  6.3017

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -43.7040     9.4916  -4.605 0.000293 ***
Air.Flow      0.8891     0.1188   7.481 1.31e-06 ***
Water.Temp    0.8166     0.3250   2.512 0.023088 *
Acid.Conc.   -0.1071     0.1245  -0.860 0.402338

---
Residual standard error: 2.569 on 16 degrees of freedom
Multiple R-Squared:  0.9488,   Adjusted R-squared:  0.9392
F-statistic: 98.82 on 3 and 16 DF,  p-value: 1.541e-10
```
normal plot now a bit skew

16 residuals show a pattern

increasing resid size with yhat
try log(y)...

> m3 <- lm(log(stack.loss) ~ Air.Flow + Water.Temp + Acid.Conc.)
> summary(m3)

Call:
  lm(formula = log(stack.loss) ~ Air.Flow + Water.Temp + Acid.Conc.)

Residuals:
   Min     1Q   Median     3Q    Max
-0.29269 -0.09734 -0.03937  0.12290  0.36558

Coefficients:  Estimate   Std. Error t value  Pr(>|t|)
(Intercept)   -0.948729    0.647721  -1.465  0.161247
Air.Flow       0.034565    0.007343   4.707  0.000203 ***
Water.Temp     0.063465    0.020038   3.167  0.005632 **
Acid.Conc.     0.002864    0.008510   0.337  0.740566

Residual standard error: 0.1766 on 17 degrees of freedom
Multiple R-Squared: 0.9033,  Adjusted R-squared: 0.8862
F-statistic: 52.92 on 3 and 17 DF,  p-value: 7.811e-09
$R^2$ is down but the diagnostic lots look pretty good...
Acid Concentration has shown negligible influence in all three models. Drop it?

Three different modeling actions under consideration:

A: Observation 21 in or out
B: y or log y as the response
C: Acid Concentration in or out
<table>
<thead>
<tr>
<th>#</th>
<th>Obs. 21</th>
<th>log acid conc.</th>
<th>$R^2$</th>
<th>SSE/df</th>
<th>Normal Plot</th>
<th>Residual versus yHat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>in</td>
<td>y in</td>
<td>0.91</td>
<td>10.5</td>
<td>21 low</td>
<td>1,2,3,4 outside</td>
</tr>
<tr>
<td>2</td>
<td>out</td>
<td>y in</td>
<td>0.95</td>
<td>6.6</td>
<td>Curved</td>
<td>1,2,3,4 outside</td>
</tr>
<tr>
<td>3</td>
<td>in</td>
<td>log y in</td>
<td>0.90</td>
<td>0.0059</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>4</td>
<td>in</td>
<td>y out</td>
<td>0.91</td>
<td>10.5</td>
<td>21 low</td>
<td>1,2,3,4 outside</td>
</tr>
<tr>
<td>5</td>
<td>out</td>
<td>y out</td>
<td>0.95</td>
<td>6.5</td>
<td>Curved</td>
<td>1,2,3,4 outside</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>log y out</td>
<td>0.90</td>
<td>0.0056</td>
<td>21 high</td>
<td>OK</td>
</tr>
<tr>
<td>7</td>
<td>out</td>
<td>log y in</td>
<td>0.92</td>
<td>0.0048</td>
<td>4 high</td>
<td>OK</td>
</tr>
<tr>
<td>8</td>
<td>out</td>
<td>log y out</td>
<td>0.92</td>
<td>0.0046</td>
<td>4 high</td>
<td>OK</td>
</tr>
</tbody>
</table>

seems clear we should drop #21 and acid conc.
if log(y) is the "correct" response, the error should increase linearly with y. Let's check this…

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>62</td>
<td>22</td>
<td>87</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>62</td>
<td>23</td>
<td>87</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>62</td>
<td>24</td>
<td>93</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>62</td>
<td>24</td>
<td>93</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>58</td>
<td>23</td>
<td>87</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>58</td>
<td>18</td>
<td>80</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>58</td>
<td>18</td>
<td>89</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>58</td>
<td>17</td>
<td>88</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>58</td>
<td>18</td>
<td>82</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>58</td>
<td>19</td>
<td>93</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>18</td>
<td>89</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>50</td>
<td>18</td>
<td>86</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>50</td>
<td>19</td>
<td>72</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>19</td>
<td>79</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>20</td>
<td>80</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

"near replicates"
<table>
<thead>
<tr>
<th>Observations</th>
<th>RSS</th>
<th>df</th>
<th>MSE</th>
<th>s</th>
<th>Yhat</th>
<th>s/Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,6,7,8</td>
<td>2.8</td>
<td>3</td>
<td>0.93</td>
<td>1.0</td>
<td>21</td>
<td>5</td>
</tr>
<tr>
<td>10, 11,12, 13,14</td>
<td>6.8</td>
<td>4</td>
<td>1.70</td>
<td>1.3</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>15,16, 17,18, 19</td>
<td>2.0</td>
<td>4</td>
<td>0.50</td>
<td>0.7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Pooled</td>
<td>11.6</td>
<td>11</td>
<td>1.05</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

no evidence that error increase with y

Also, MSE here is much smaller than, e.g., model 5 (MSE=6.5) Something is not right! (can use an F-test here)
it seems that when air flow exceeds 60, the plant takes about a day to come to equilibrium

"line-out" is the term used by plant operators

This would suggest permanently dropping points 1, 3, 4, and 21 since they correspond to transient states

Now revisit the log y and acid concentration issues…
1, 3, 4, and 21 removed

<table>
<thead>
<tr>
<th>#</th>
<th>Obs. 21</th>
<th>log</th>
<th>acid conc.</th>
<th>$R^2$</th>
<th>SSE/df</th>
<th>Normal Plot</th>
<th>Residual versus yHat</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>out</td>
<td>y</td>
<td>in</td>
<td>0.975</td>
<td>1.6</td>
<td>20 low</td>
<td>curvature?</td>
</tr>
<tr>
<td>10</td>
<td>out</td>
<td>y</td>
<td>out</td>
<td>0.973</td>
<td>1.6</td>
<td>OK</td>
<td>curvature?</td>
</tr>
<tr>
<td>11</td>
<td>out</td>
<td>log y</td>
<td>in</td>
<td>0.92</td>
<td>0.0032</td>
<td>20 high</td>
<td>curvature?</td>
</tr>
<tr>
<td>12</td>
<td>out</td>
<td>log y</td>
<td>out</td>
<td>0.92</td>
<td>0.0031</td>
<td>20 high</td>
<td>curvature?</td>
</tr>
</tbody>
</table>
model 9
drop acid conc. again
try adding airflow$^2$

\begin{verbatim}
lm(formula = stack.loss ~ Air.Flow + Water.Temp + I(Air.Flow^2),
    subset = c(-1, -3, -4, -21))

Residuals:
   Min     1Q Median     3Q    Max
-2.0177 -0.6530 -0.1252  0.5101  2.3429

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)         -15.409290  12.602668  -1.223  0.24315
Air.Flow            -0.069142   0.398419  -0.174  0.86490
Water.Temp          0.527804   0.150079   3.517  0.00379 **
I(Air.Flow^2)       0.006818   0.003178   2.145  0.05139 .

Residual standard error: 1.125 on 13 degrees of freedom
Multiple R-Squared: 0.9799,    Adjusted R-squared: 0.9752
F-statistic: 210.8 on 3 and 13 DF,  p-value: 2.854e-11
\end{verbatim}
probably makes no sense to go further

Could try a factorial study of $x_1^2$, $x_2^2$, $x_1x_2$, log y, etc.

SSE/df is now 1.26 compared with the "minimum" 1.05
model selection in linear regression

basic problem: how to choose between competing linear regression models

model too small: "underfit" the data; poor predictions; high bias; low variance

model too big: "overfit" the data; poor predictions; low bias; high variance

model just right: balance bias and variance to get good predictions
Bias-Variance Tradeoff

High Bias - Low Variance
Score function should embody the compromise

Low Bias - High Variance
“overfitting” - modeling the random component
model selection in regression has two facets:

1. assign a score to each model

2. search for models with good scores
linear regression model scores

consider the problem of selecting a "good" subset of $k$ candidate predictors

$$S \subseteq \{1, \ldots, k\}$$

$$\chi_S = \{X_j : j \in S\}$$

true coefficients

least squares estimates

design matrix

estimated regression function

$$\hat{Y}_i(S) = \hat{r}_S(X_i)$$
prediction risk:

\[
R(S) = \sum_{i=1}^{n} E_{y,y^*}(\hat{Y}_i(S) - Y_i^*)^2
\]

value of future observation at \(X_i\)

goal: pick the model that minimizes \(R(S)\)

training error:

\[
\hat{R}_{TR}(S) = \sum_{i=1}^{n} \left(\hat{Y}_i(S) - Y_i\right)^2
\]

bad estimate of risk!
Theorem: The training error is a downward-biased estimate of the prediction risk:

$$E\left(\hat{R}_{TR}(S)\right) < R(S)$$

bias\left(\hat{R}_{TR}(S)\right) = E_y\left(\hat{R}_{TR}(S)\right) - R(S) = -2 \sum_{i=1}^{n} \text{Cov}\left(\hat{Y}_i, Y_i\right)$$

for linear models with $|S|$ predictors:

$$\sum_{i=1}^{n} \text{Cov}\left(\hat{Y}_i, Y_i\right) = |S|\sigma^2_{\epsilon}$$

tends to be large when the model is large
The obvious thing to do is estimate the bias and adjust!

\[ C_p = \frac{\hat{R}_{TR}(S)}{2|S|\hat{\sigma}_\varepsilon^2} + \frac{2\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2} \]

- \( \hat{R}_{TR}(S) \): how well the model fits the training data; smaller is better
- \( 2|S|\hat{\sigma}_\varepsilon^2 \): complexity penalty; bigger model, bigger penalty

"Mallows \( C_p \) statistic"
Akaike Information Criterion is one alternative:

\[ AIC = l_S - |S| \]

where \( l_S \) is the maximized log-likelihood

(very similar to \( C_p \) in normal linear regression models)

- can use cross-validation to estimate prediction risk

- for linear regression, there are short cut formulae that can compute the CV estimate from a single (full) model fit
AIC in R is multiplied by -2 (so smaller is better)

$AIC_R = -2l_s + 2|S|$

for linear regression with normal errors, the log likelihood is:

$-\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|y - X\beta\|^2$

plugging the MLE for $\beta$:

$-\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \text{RSS}$
Thus, if $\sigma$ is known:

$$AIC = n \log 2\pi + n \log \sigma^2 + \frac{1}{\sigma^2} \text{RSS} + 2|S|$$

constant

If $\sigma$ is unknown:

$$AIC = n \log (\text{RSS}/n) + 2|S| + \text{const}$$
Bayesian Information Criterion is one alternative:

\[ \text{BIC} = l_S - \frac{|S|}{2} \log n \]

where \( l_S \) is the maximized log-likelihood

Bayesian interpretation: suitably normalized, BIC scores can be interpreted as approximate posterior model probabilities: \( P(S_j \mid \text{Data}) \)
Bayesian Criterion

\[ p(M_k \mid D) \propto p(D \mid M_k)p(M_k) \]

\[ = p(M_k) \int p(D \mid \theta_k, M_k)p(\theta_k \mid M_k)d\theta_k \]

• Typically impossible to compute analytically
• All sorts of Monte Carlo approximations
Suppose $M_0$ simplifies $M_1$ by setting one parameter (say $q_1$) to some constant (typically zero)

If $p_1(q_2 \mid q_1 = 0) = p_0(q_2)$ then:

$$\frac{p(\text{data} \mid M_0)}{p(\text{data} \mid M_1)} = \frac{p(q_1 = 0 \mid M_1, \text{data})}{p(q_1 = 0 \mid M_1)}$$

**Savage-Dickey Density Ratio**
Laplace Method for $p(D|M)$

let $l(\theta) = \frac{\log(L(\theta))}{n} + \frac{\log p(\theta)}{n}$

(i.e., the log of the integrand divided by $n$)

then $p(D) = \int e^{nl(\theta)} d\theta$

Laplace’s Method:

$p(D) \approx \int \exp[nl(\tilde{\theta}) - n(\theta - \tilde{\theta})^2/(2\sigma^2)]d\theta$

where $\sigma^2 = -1/l'''(\tilde{\theta})$ and $\tilde{\theta}$ is the posterior mode
Laplace cont.

\[ p(D) = \int \exp [nl(\tilde{\theta}) - n(\theta - \tilde{\theta})^2 / (2\sigma^2)] d\theta \]

\[ \approx \sqrt{2\pi} \sigma n^{-1/2} \exp \{nl(\tilde{\theta})\} \]

• Tierney & Kadane (1986, JASA) show the approximation is O(n^{-1})

• Using the MLE instead of the posterior mode is also O(n^{-1})

• Using the expected information matrix in s is O(n^{-1/2}) but convenient since often computed by standard software

• Raftery (1993) suggested approximating \( \tilde{\theta} \) by a single Newton step starting at the MLE

• Note the prior is explicit in these approximations
Monte Carlo Estimates of $p(D|M)$

$$p(D) = \int p(D | \theta) p(\theta) d\theta$$

Draw iid $\theta_1, \ldots, \theta_m$ from $p(\theta)$:

$$\hat{p}(D) = \frac{1}{m} \sum_{i=1}^{m} p(D | \theta^{(i)})$$

In practice has large variance
Monte Carlo Estimates of $p(D|M)$ (cont.)

Draw iid $\theta_1, \ldots, \theta_m$ from $p(\theta|D)$:

$$\hat{p}(D) = \frac{\sum_{i=1}^{m} w_i p(D | \theta^{(i)})}{\sum_{i=1}^{m} w_i}$$

$w_i = \frac{p(\theta^{(i)})}{p(\theta^{(i)} | D)} = \frac{p(\theta^{(i)}) p(D)}{p(D | \theta^{(i)}) p(\theta^{(i)})}$

“Importance Sampling”
Monte Carlo Estimates of $p(D|M)$ (cont.)

$$\hat{p}(D) = \frac{\sum_{i=1}^{m} \frac{p(D)}{p(D | \theta^{(i)})} p(D | \theta^{(i)})}{\sum_{i=1}^{m} \frac{p(D)}{p(D | \theta^{(i)})}}$$

$$= \left\{ \frac{1}{m} \sum_{i=1}^{m} p(D | \theta^{(i)})^{-1} \right\}^{-1}$$

Newton and Raftery’s “Harmonic Mean Estimator”

Unstable in practice and needs modification
\[ p(D|M) \] from Gibbs Sampler Output

First note the following identity (for any \( \theta^* \)):

\[
p(D) = \frac{p(D | \theta^*) p(\theta^*)}{p(\theta^* | D)}
\]

\( p(D | \theta^*) \) and \( p(\theta^*) \) are usually easy to evaluate.

What about \( p(\theta^* | D) \)?

Suppose we decompose \( \theta \) into \( (\theta_1, \theta_2) \) such that \( p(\theta_1 | D, \theta_2) \) and \( p(\theta_2 | D, \theta_1) \) are available in closed-form...

Chib (1995)
$p(D|M)$ from Gibbs Sampler

Output

$p(\theta_1^*, \theta_2^* \mid D) = p(\theta_2^* \mid D, \theta_1^*) p(\theta_1^* \mid D)$

The Gibbs sampler gives (dependent) draws from $p(\theta_1, \theta_2 \mid D)$ and hence marginally from $p(\theta_2 \mid D)$…

$p(\theta_1^* \mid D) = \int p(\theta_1^* \mid D, \theta_2) p(\theta_2 \mid D) d\theta_2$

$\approx \frac{1}{G} \sum_{g=1}^{G} p(\theta_1^* \mid D, \theta_2^{(g)})$

“Rao-Blackwellization”
Bayesian Information Criterion

\[ S_{BIC}(M_k) = -S_L(\hat{\theta}_k; M_k) - \frac{d_k}{2} \log n, \quad k = 1, \ldots, K \]

\((S_L\) is the negative log-likelihood)

• BIC is an \(O(1)\) approximation to \(p(D|M)\)

• Circumvents explicit prior

• Approximation is \(O(n^{-1/2})\) for a class of priors called “unit information priors.”

• No free lunch (Weakliem (1998) example)
Srole (1956): "It's hardly fair to bring a child into the world now the way things look for the future." The data are from the 1993-94 General Social Survey; respondents were given the options of agreeing or disagreeing, and the few who could not choose are excluded from the analysis. The sample of 2,266 valid responses is composed of 44.0

\[ \log(n_{ij}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \Theta x_3. \]  \hspace{1cm} (4)

In this parameterization, \( x_1 \) is a dummy variable that is zero in row 1 and one in row 2, \( x_2 \) is a dummy variable that is zero in column 1 and one in column 2, and \( x_3 \) is a dummy variable that is one in column 2 of row 2 and zero otherwise. The maximum likelihood estimate of \( \Theta \) is the logarithm of the observed odds ratio \( (n_{11}n_{22})/(n_{12}n_{21}) \), so another way to put the question is to ask if \( \Theta \) is equal to zero.

The sample contains 412 men who agree with the statement, 583 men who disagree, 584 women who agree, and 687 women who disagree. The \( L^2 \) for the model of independence is 4.68 with one degree of
Figure 1: Bayes Factors for Model of No Association in Anemia by Gender Table:
    Normal Prior Distribution With Mean Zero

about 4.0 (the exact figure is 4.07). With this prior distribution, the 95
percent range for possible values of the odds ratio would extend from
1/2,914 to 2,914, whereas the 50 percent range would extend from
1/14.7 to 14.7. In other words, adopting this prior distribution is
equivalent to saying that if there is any association between the vari-
able, there is a 50 percent chance that the absolute value of the odds
ratio will be more than 14.7 or less than 1/14.7. As discussed above,
Deviance Information Criterion

• Deviance is a standard measure of model fit:
  \[ D(y, \theta) = -2 \log p(y | \theta) \]

• Can summarize in two ways… at posterior mean or mode:
  
  (1) \[ D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y)) \]
  
  or by averaging over the posterior:

  (2) \[ D_{avg}(y) = E(D(y, \theta) | y) \]

(2) will be bigger (i.e., worse) than (1)
Deviance Information Criterion

\[ p_D^{(1)} = D_{avg}(y) - D_{\hat{\theta}}(y) \]

is a measure of model complexity.

• In the normal linear model \( p_D^{(1)} \) equals the number of parameters.

• More generally \( p_D^{(1)} \) equals the number of unconstrained parameters.

• DIC = \( D_{avg}(y) + p_D^{(1)} \)

• Approximately equal to \( E[D(y^{rep}, \hat{\theta}(y))] \)
model search

- forward stepwise
- backward stepwise
- all-subsets
- genetic algorithms
- stochastic search

start with the empty model and add one variable at a time greedily

start with the full model and delete one variable at a time greedily

stepwise methods can get stuck at local modes
Zheng-Loh

1. fit the full model with all $d$ predictors and let:

   \[ W_j = \frac{\hat{\beta}_j}{s\hat{\varepsilon}(\hat{\beta}_j)} \]

2. Order the statistics in absolute value from largest to smallest:

   \[ |W_{(1)}| \geq |W_{(2)}| \geq \cdots \geq |W_{(d)}| \]

3. Let $\hat{j}$ be the value of $j$ that minimizes

   \[ \text{RSS}(j) + j\hat{\sigma}^2 \log n \]

   \( \hat{\sigma}^2 \) is the variance estimate from the full model
   \( \text{RSS}(j) \) is from the model using \( x_{(1)}, \ldots, x_{(j)} \)

4. Choose as the final model, the regression with the terms with the largest $W$'s
Computing: Variable Selection via Stepwise Methods

- Efroymson’s 1960 algorithm still the most widely used

1. Enter into the (linear) regression model any variables that are to be “forced in.”
2. Find the variable from those not in the model but available for inclusion that has the largest $F$-to-enter value. If it is at least as great as a prespecified value, $F_{in}$, then add the variable to the model. Stop if no variable can be added.
3. Find that variable among those in the model, other than those forced in, that has the smallest $F$-to-remove value. If it is less than a prespecified value, $F_{out}$, then drop the variable from the model. Repeat this step until no further variables can be dropped; then go to step 2.
Efroymson

- F-to-Enter
  \[ \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n - p - 1)} \}
  Distribution not even remotely like \( F \)

- F-to-Remove
  \[ \frac{RSS_{p-1} - RSS_p}{RSS_p/(n - p)} \}

- Guaranteed to converge

- Not guaranteed to converge to the right model...
Trouble

**An artificial data set**

<table>
<thead>
<tr>
<th>Observation number</th>
<th>Predictors</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X_1$</td>
<td>$X_2$</td>
<td>$X_3$</td>
<td>$Y$</td>
</tr>
<tr>
<td>1</td>
<td>1000</td>
<td>1002</td>
<td>0</td>
<td>-2</td>
</tr>
<tr>
<td>2</td>
<td>-1000</td>
<td>-999</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>-1000</td>
<td>-1001</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1000</td>
<td>998</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

- $Y = X_1 - X_2$
- $Y$ almost orthogonal to $X_1$ and $X_2$
- Forward selection and Efroymson pick $X_3$ alone
More Trouble

- Berk Example with 4 variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Highest $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>0.01</td>
</tr>
<tr>
<td>$X_2, X_3$</td>
<td>0.99</td>
</tr>
<tr>
<td>$X_1, X_2, X_4$</td>
<td>0.994</td>
</tr>
</tbody>
</table>

- The forward and backward sequence is $(X_1, X_1X_2, X_1X_2X_4)$

- The $R^2$ for $X_1X_2$ is 0.015
“Detroit” example, \( N=13, d=11 \)

First variable selected in forward selection is the first variable eliminated by backward elimination

Best subset of size 3 gives RSS of 6.8

Forward’s best set of 3 has RSS = 21.2; Backward’s gets 23.5
Variable selection with pure noise using leaps

```r
y <- rnorm(100)
xx <- matrix(rnorm(4000),ncol=40)
dimnames(xx) <- list(NULL,paste("X",1:40,sep=""))

library(leaps)
xx.subsets <- regsubsets(xx, y, method="exhaustive", nvmax=3, nbest=1)
subvar <- summary(xx.subsets)$which[3,-1]
best3.lm <- lm(y ~ -1 + xx[, subvar])
print(summary(best3.lm, corr=FALSE))

or...bestsetNoise(m=100,n=40)
```
run this experiment ten times:

- all three significant at $p<0.01$ 1
- all three significant at $p<0.05$ 3
- two out of three significant at $p<0.05$ 3
- one out of three significant at $p<0.05$ 1
The Impact of Model Selection on Inference in Linear Regression

CLIFFORD M. HURVICH and CHIH–LING TSAI*

Table 2. Coverage Rates for Confidence Regions With Nominal Rate 1 – \( \alpha \)

\( (n = 30, p_c = 4, \text{model order } p \text{ chosen by } \text{AIC}) \)

<table>
<thead>
<tr>
<th>( p )</th>
<th>( 1 - \alpha = .9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.000 (0/15)</td>
</tr>
<tr>
<td>4</td>
<td>.867 (263/303)</td>
</tr>
<tr>
<td>5</td>
<td>.838 (62/74)</td>
</tr>
<tr>
<td>6</td>
<td>.660 (35/53)</td>
</tr>
<tr>
<td>7</td>
<td>.600 (33/55)</td>
</tr>
<tr>
<td>OCR</td>
<td>.786 (393/500)</td>
</tr>
</tbody>
</table>

Table 5. Coverage Rates for Confidence Regions With Nominal Rate 1 – \( \alpha \)

\( (n = 30, p_c = 4, \text{model order } p \text{ chosen by } \text{BIC}) \)

<table>
<thead>
<tr>
<th>( p )</th>
<th>( 1 - \alpha = .9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.000 (0/62)</td>
</tr>
<tr>
<td>4</td>
<td>.890 (331/372)</td>
</tr>
<tr>
<td>5</td>
<td>.833 (30/36)</td>
</tr>
<tr>
<td>6</td>
<td>.313 (5/16)</td>
</tr>
<tr>
<td>7</td>
<td>.286 (4/14)</td>
</tr>
<tr>
<td>OCR</td>
<td>.740 (370/500)</td>
</tr>
</tbody>
</table>
Bayesian Model Averaging

- If we believe that one of the candidate models generated the data, then the predictively optimal strategy is to average over all the models.

- If $Q$ is the inferential target, Bayesian Model Averaging (BMA) computes:

\[
p(Q) = \sum_{i \in I} p(Q|D, M_i)p(M_i|D)
\]

- Substantial empirical evidence that BMA provides better prediction than model selection