STAT 545

LINEAR MODELS

Conditional expectation linear in parameters.

$$\begin{aligned}
E(Y|X_1, X_2) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\
E(Y|X_1, X_2) &= \beta_0 + \beta_1 X_1^3 + \beta_2 X_1 X_2 \\
E(Y|X_1, X_2) &= \beta_0 + \beta_1 I_{\{X_1>0\}} \\
E(Y|X_1, X_2) &= \beta_0 + \beta_1 \exp(X_1) + \beta_2 X_2
\end{aligned}$$

But not

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \exp(-\beta_2 X_2)$$

Notation

Repeated realizations of $(Y, X_1, \ldots, X_p)$, where

$$Y|X_1, \ldots, X_p \sim N\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \sigma^2\right)$$

Or $i = 1, \ldots, n$ indexes observations, $j = 1, \ldots, p$ indexes predictors, observe vector of responses Y (entries Y_i) and design matrix X (entries X_{ij}).

$$Y|X \sim N_n(X\beta, \sigma^2 I_n).$$

ML/LS estimator: $\hat{\beta} = \text{argmin}_\beta \|Y - X\beta\|^2 = (X^T X)^{-1} X^T Y$.

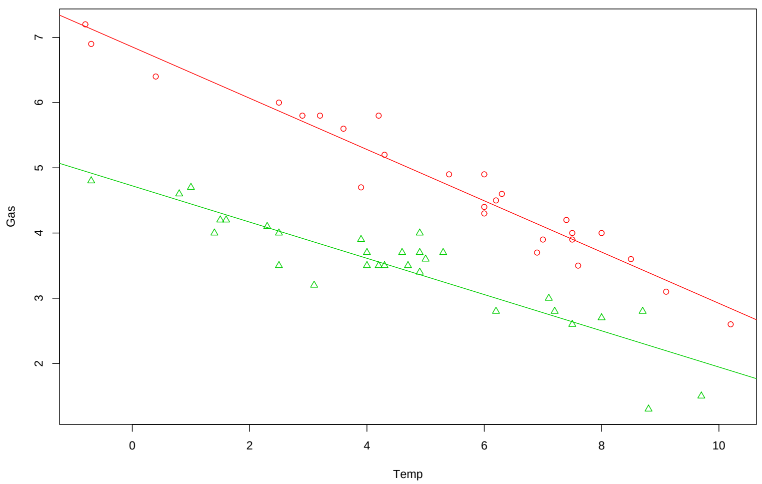
NOTE UBIQUITY OF DESIGN MATRIX, RESPONSE VECTOR FORMULATION: linear regression, multiple linear regression, ANOVA, curve-fitting,....

Software: lm() function.

```
> tmp <- lm(y~x)
> coef(tmp)     ### or tmp$coef
> resid(tmp)    ### or tmp$resid
> fitted(tmp)   ### or tmp$fitted
```

Simple Example

```
> attach(whiteside)
> plot(Temp, Gas, pch=as.numeric(Insul),
               col=1+as.numeric(Insul))
> tmp1 <- lm(Gas~Temp, data=whiteside,
          subset=Insul=="Before")
> abline(tmp1, col=2)
...
> names(tmp1)
 [1] "coefficients" "residuals"     "effects"
 [4] "rank"         "fitted.values" "assign"
 [7] "qr"           "df.residual"   "xlevels"
[10] "call"         "terms"         "model"
```



```
> summary(tmp1)
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = ...
Residuals:
     Min       1Q   Median       3Q      Max
-0.62020 -0.19947  0.06068  0.16770  0.59778
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.85383    0.11842   57.88   <2e-16 ***
Temp        -0.39324    0.01959  -20.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ...

Residual standard error: 0.2813 on 24 degrees of freedom
Multiple R-Squared: 0.9438,     Adjusted R-squared: 0.9415
F-statistic: 403.1 on 1 and 24 DF,  p-value: < 2.2e-16
```

Model Formulae

```
fake.data <- cbind(
  data.frame("y"=rnorm(54)), data.frame("x1"=rnorm(54)),
  data.frame("x2"=rnorm(54)),
  data.frame("a"=factor(c(rep("ubc",18),rep("sfu",18),
    rep("vic",18)), levels=c("ubc","sfu","vic"))),
  data.frame("b"=ordered(rep(c(rep("sm",3),rep("md",3),
    rep("lg",3)),6), levels=c("sm","md","lg"))),
  data.frame("c"=factor(rep(c("rd","gr","bl"),18),
                        levels=c("rd","gr","bl"))) )
```

```
> print(fake.data)
        y     x1     x2   a  b  c
1  -0.288 -0.109  1.486 ubc sm rd
2  -0.397 -0.862  0.905 ubc sm gr
3   0.179 -1.482 -1.450 ubc sm bl
4  -0.863 -0.457 -0.603 ubc md rd
5   0.926 -0.258  0.733 ubc md gr
6  -0.594  0.739 -0.413 ubc md bl
7   0.729  0.704 -0.384 ubc lg rd
8  -0.130  1.661  0.134 ubc lg gr
9   1.734 -1.010 -0.464 ubc lg bl
10  2.012 -0.469  1.085 ubc sm rd
...
52  0.416 -1.903 -0.113 vic lg rd
53  0.579 -1.294  0.975 vic lg gr
54  0.567  1.380 -0.953 vic lg bl
```

```
> opt <- lm(y ~ x1 +x2, data=fake.data)
> summary(opt)

Call:
lm(formula = y ~ x1 + x2, data = fake.data)
Coefficients:
(Intercept)           x1           x2
      0.020        0.154        0.018

Call:
lm(formula = y ~ -1 + x1 + x2, data = fake.data)
Coefficients:
   x1     x2
0.158  0.017
```

```
Call:
lm(formula = y ~ x1 * x2, data = fake.data)
Coefficients:
(Intercept)           x1           x2        x1:x2
      0.040        0.119        0.042       -0.194

Call:
lm(formula = y ~ x1 + x2 + I(x1 * x2), data = fake.data)
Coefficients:
(Intercept)           x1           x2   I(x1 * x2)
      0.040        0.119        0.042       -0.194
```

```
> getOption("contrasts")
        unordered           ordered
"contr.treatment"      "contr.poly"

>  opt <- lm(y ~ a + c, data=fake.data)
>  opt$coef
(Intercept)        asfu       auvic        cgr       cbl
      0.245      -0.359      -0.209      0.044    -0.275

> dummy.coef(opt)
Full coefficients are
(Intercept):      0.24
a:                 ubc    sfu   uvic
                  0.00  -0.36  -0.21
c:                  rd     gr     bl
                 0.000  0.044 -0.275
```

```
> options(constrasts= c("contr.sum", "contr.poly"))

>  opt <- lm(y ~ a + c, data=fake.data)
>  opt$coef
(Intercept)        a1        a2        c1        c2
     -0.022     0.189    -0.170     0.077     0.121

> dummy.coef(opt)
Full coefficients are
(Intercept):    -0.022
a:                 ubc    sfu   uvic
                 0.189 -0.170 -0.020
c:                  rd     gr     bl
                 0.077  0.121 -0.198
```

```
lm(formula = y ~ ., data = fake.data)

Coefficients:
(Intercept)           x1           x2         asfu
    0.02083      0.14795      0.00039      0.14433
       avic          bmd          blg          cgr
    0.12052     -0.08310      0.21444     -0.28178
        cbl
   -0.11705
```

```
lm(formula = y ~ x1 + x2 + a * b, data = fake.data)
Coefficients:
(Intercept)           x1           x2         asfu
    -0.2077       0.1461       0.0048       0.1222
       avic          bmd          blg     asfu:bmd
     0.4268       0.2538       0.1646      -0.0966
   avic:bmd     asfu:blg     avic:blg
    -0.9073       0.1664      -0.0115

lm(formula = y ~ x1 + b/x2, data = fake.data)
Coefficients:
(Intercept)           x1          bmd          blg
     -0.073        0.157       -0.081        0.266
     bsm:x2       bmd:x2       blg:x2
      0.181       -0.227        0.019
```

```
> opt <- lm(y ~ a * b * c, data = fake.data)

> names(opt$coef)
 [1] "(Intercept)"  "asfu"         "avic"
 [4] "bmd"          "blg"          "cgr"
 [7] "cbl"          "asfu:bmd"     "avic:bmd"
[10] "asfu:blg"     "avic:blg"     "asfu:cgr"
[13] "avic:cgr"     "asfu:cbl"     "avic:cbl"
[16] "bmd:cgr"      "blg:cgr"      "bmd:cbl"
[19] "blg:cbl"      "asfu:bmd:cgr" "avic:bmd:cgr"
[22] "asfu:blg:cgr" "avic:blg:cgr" "asfu:bmd:cbl"
[25] "avic:bmd:cbl" "asfu:blg:cbl" "avic:blg:cbl"
```