Multiple Regression Model (Chap 6)
Basic tools for building regression models: indicator variables, splines, interactions

Yen-Yi Ho

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

More than one predictor

Data

	Y	X	Z
1	0.72	0.37	0
2	0.65	0.19	0
3	0.81	0.11	0
4	-0.06	-0.44	0
5	1.39	-0.31	0
6	-0.04	-0.39	1
7	-0.09	-0.20	1
8	-0.31	-0.23	1
9	0.85	-0.01	1
10	0.35	-0.45	1
...			

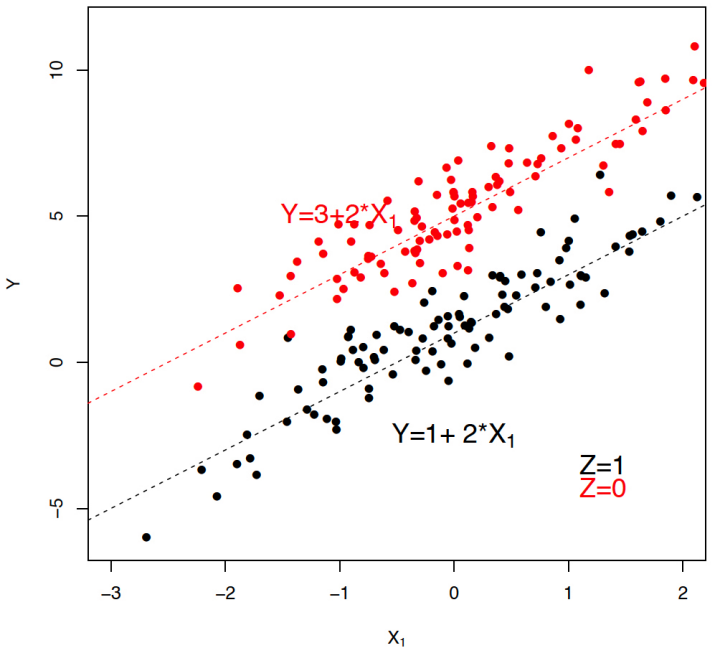
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$$

In other words (or, equations):

$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \epsilon_i, & \text{if } Z_i = 0 \\ (\beta_0 + \beta_2) + \beta_1 X_i + \epsilon_i, & \text{if } Z_i = 1 \end{cases}$$

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 Z_i + \epsilon_i$$



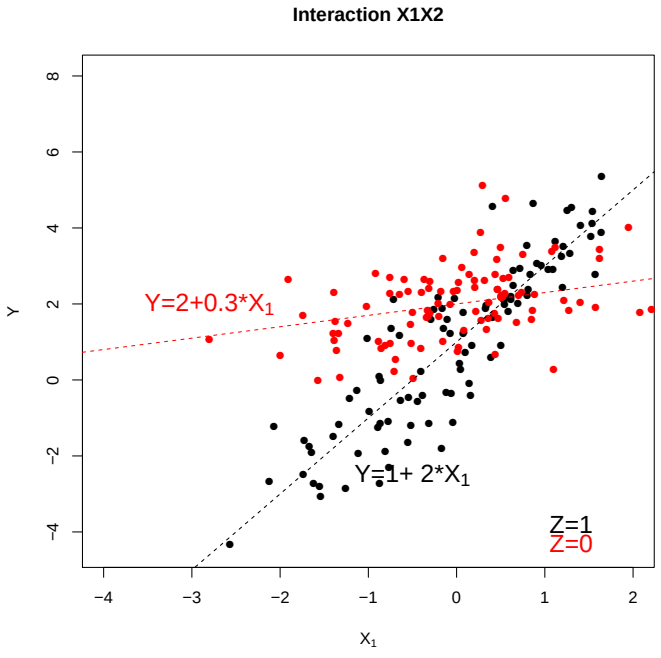
$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \epsilon_i, & \text{if } Z_i = 0 \\ (\beta_0 + \beta_2) + \beta_1 X_i + \epsilon_i, & \text{if } Z_i = 1 \end{cases}$$

$\rightarrow$ Assuming the same slope for both $Z = 0$ and $Z = 1$.

Multiple Linear Regression: Interaction

When slopes are different in $Z = 0$ vs. $Z = 1$,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + \epsilon_i$$



$$Y_i = \begin{cases} \beta_0 + \beta_1 X_i + \epsilon_i, & \text{if } Z_i = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i + \epsilon_i, & \text{if } Z_i = 1 \end{cases}$$

Effect modification: the effect of X differs depending on the level of Z.

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

- Gender: 0: women, 1:men
- Smoke: 0: No, 1:Yes

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.97623    0.09036  110.41   <2e-16 ***
gendermale         10.07630    0.12399   81.27   <2e-16 ***
smokeYes           -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the predicted average income for a man who smoke?

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.97623    0.09036  110.41   <2e-16 ***
gendermale          10.07630    0.12399   81.27   <2e-16 ***
smokeYes            -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the predicted average income for a man who smoke?
  $b_0 + b_1 + b2 + b_3 = 9.98 + 10.08 - 4.06 + 2.01 = \$18.01$
- What is the predicted average income for a woman who smoke?

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.97623    0.09036  110.41   <2e-16 ***
gendermale          10.07630    0.12399   81.27   <2e-16 ***
smokeYes            -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the predicted average income for a man who smoke?
  $b_0 + b_1 + b2 + b_3 = 9.98 + 10.08 - 4.06 + 2.01 = \$18.01$
- What is the predicted average income for a woman who smoke?
  $b_0 + b2 = 9.98 - 4.06 = \$5.92$

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.97623    0.09036  110.41   <2e-16 ***
gendermale          10.07630    0.12399   81.27   <2e-16 ***
smokeYes            -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the predicted average income for a man who smoke?
  $b_0 + b_1 + b2 + b_3 = 9.98 + 10.08 - 4.06 + 2.01 = \$18.01$
- What is the predicted average income for a woman who smoke?
  $b_0 + b2 = 9.98 - 4.06 = \$5.92$
- What is the predicted difference in income between men who smoke and who don't?

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.97623    0.09036  110.41   <2e-16 ***
gendermale          10.07630    0.12399   81.27   <2e-16 ***
smokeYes            -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the predicted average income for a man who smoke?
  $b_0 + b_1 + b2 + b_3 = 9.98 + 10.08 - 4.06 + 2.01 = \$18.01$
- What is the predicted average income for a woman who smoke?
  $b_0 + b2 = 9.98 - 4.06 = \$5.92$
- What is the predicted difference in income between men who smoke and who don't?
  $(b_0 + b_1 + b_2 + b_3) - (b_0 + b_1) = b_2 + b_3$
- What is the predicted difference in income between women who smoke and who don't?

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          9.97623    0.09036  110.41   <2e-16 ***
gendermale          10.07630    0.12399   81.27   <2e-16 ***
smokeYes            -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the predicted average income for a man who smoke?
  $b_0 + b_1 + b2 + b_3 = 9.98 + 10.08 - 4.06 + 2.01 = \$18.01$
- What is the predicted average income for a woman who smoke?
  $b_0 + b2 = 9.98 - 4.06 = \$5.92$
- What is the predicted difference in income between men who smoke and who don't?
  $(b_0 + b_1 + b_2 + b_3) - (b_0 + b_1) = b_2 + b_3$
- What is the predicted difference in income between women who smoke and who don't?
  b_2

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.97623    0.09036  110.41   <2e-16 ***
gendermale         10.07630    0.12399   81.27   <2e-16 ***
smokeYes           -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the interpretation of b_3?
  The difference of the difference in average income between men who smoke and who don't and between women who smoke and who don't.

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.97623    0.09036  110.41   <2e-16 ***
gendermale         10.07630    0.12399   81.27   <2e-16 ***
smokeYes           -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the interpretation of b_3?
  The difference of the difference in average income between men who smoke and who don't and between women who smoke and who don't.
- What is the interpretation of $b_1 + b_3$?

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.97623    0.09036  110.41   <2e-16 ***
gendermale         10.07630    0.12399   81.27   <2e-16 ***
smokeYes           -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the interpretation of b_3?
  The difference of the difference in average income between men who smoke and who don't and between women who smoke and who don't.
- What is the interpretation of $b_1 + b_3$?
  The difference of income between men who smoke versus women who smoke.

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.97623    0.09036  110.41   <2e-16 ***
gendermale         10.07630    0.12399   81.27   <2e-16 ***
smokeYes           -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the interpretation of b_3?
  The difference of the difference in average income between men who smoke and who don't and between women who smoke and who don't.
- What is the interpretation of $b_1 + b_3$?
  The difference of income between men who smoke versus women who smoke.
- What is the interpretation of b_1?

Income on gender and smoking, include interaction

$$E(Income) = \beta_0 + \beta_1(gender) + \beta_2(smoke) + \beta_3(gender \times smoke)$$

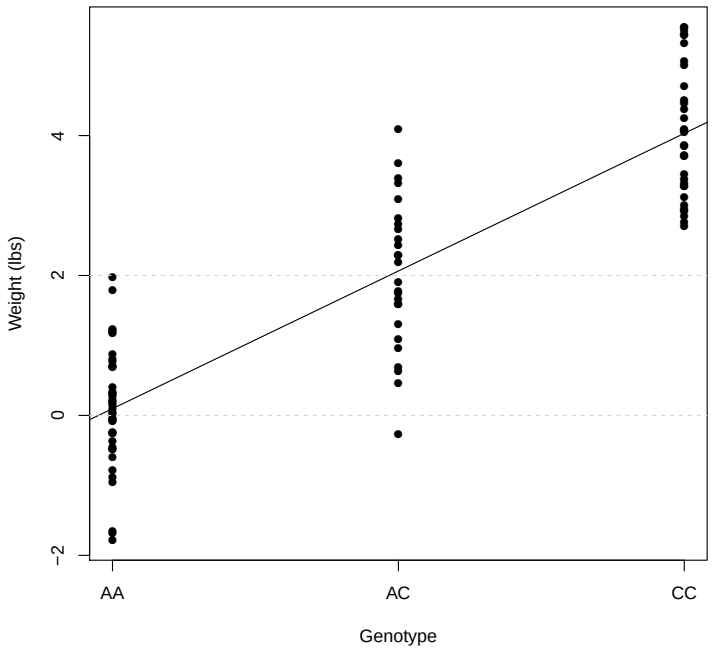
```
> fit2<-lm(income ~ gender*smoke)
> summary(fit2)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.97623    0.09036  110.41   <2e-16 ***
gendermale         10.07630    0.12399   81.27   <2e-16 ***
smokeYes           -4.05857    0.12881  -31.51   <2e-16 ***
gendermale:smokeYes  2.01454    0.17713   11.37   <2e-16 ***
```

- What is the interpretation of b_3?
  The difference of the difference in average income between men who smoke and who don't and between women who smoke and who don't.
- What is the interpretation of $b_1 + b_3$?
  The difference of income between men who smoke versus women who smoke.
- What is the interpretation of b_1?
  The difference of income between non-smoking men and non-smoking women.

Building Regression Model: Linear

$$\widehat{Y} = \beta_0 + \beta_1 X$$

	y	Genotype	x
1	2.85	CC	2
2	0.40	AA	0
3	3.28	CC	2
4	1.80	AA	0
5	2.19	CA	1
6	1.97	AA	0
7	0.64	CA	1

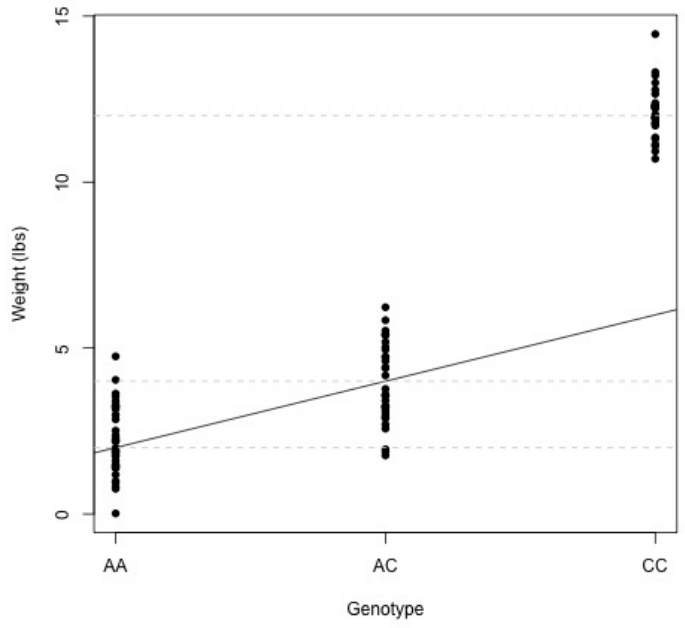



```
>fit<-lm(y ~ x)
>summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1713     0.1821   0.941    0.349
x             1.9490     0.2313   8.427 3.09e-13 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Building Regression Model

$$\widehat{Y} = \beta_0 + \beta_1 X$$

	y	Genotype	x
1	2.85	CC	?
2	0.40	AA	?
3	3.28	CC	?
4	1.80	AA	?
5	2.19	CA	?
6	1.97	AA	?
7	0.64	CA	?

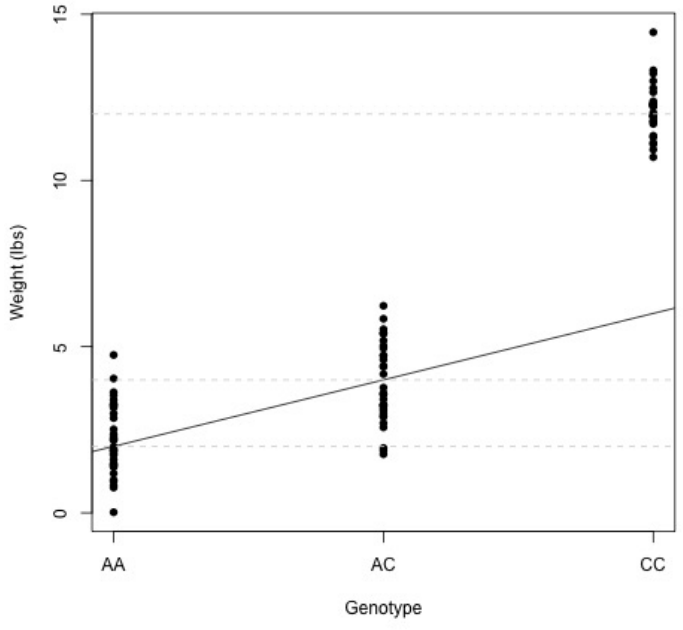


Building Regression Model: Dummy Variables

$$\widehat{Y} = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2$$

$$\widehat{Y} = \beta_0 + \beta_1 \times I(genotype = CA) + \beta_2 \times I(genotype = CC)$$

	y	Genotype	x1	x2
1	2.85	CC	0	1
2	0.40	AA	0	0
3	3.28	CC	0	1
4	1.80	AA	0	0
5	2.19	CA	1	0
6	1.97	AA	0	0
7	0.64	CA	1	0



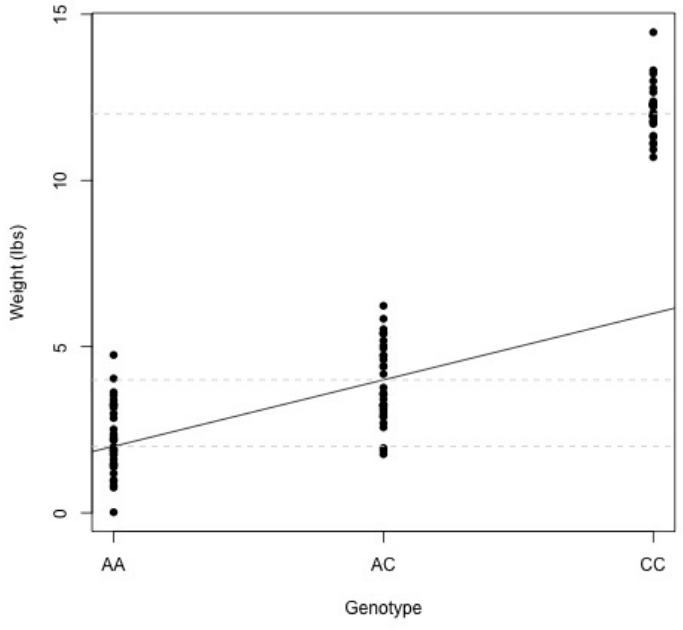

```
>fit<-lm(y ~ x1 + x2)
>summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2065     0.1694  13.024  < 2e-16 ***
x1            1.6969     0.2448   6.931 4.62e-10 ***
x2            9.9155     0.2556  38.796  < 2e-16 ***
```

Building Regression Model: Dummy Variables

$$\widehat{Y} = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2$$

$$\widehat{Y} = \beta_0 + \beta_1 \times I(genotype = CA) + \beta_2 \times I(genotype = CC)$$

	y	Genotype	x1	x2
1	2.85	CC	0	1
2	0.40	AA	0	0
3	3.28	CC	0	1
4	1.80	AA	0	0
5	2.19	CA	1	0
6	1.97	AA	0	0
7	0.64	CA	1	0




```
>fit<-lm(y ~ x1 + x2)
>summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2065     0.1694  13.024  < 2e-16 ***
x1            1.6969     0.2448   6.931 4.62e-10 ***
x2            9.9155     0.2556  38.796  < 2e-16 ***
```

Based on the regression output, what is the difference of average weight between CA and AA? CC and AA? CC and CA?

Indicator Variables

General case: Use p-1 indicator ("dummy") variables to represent p groups

Group	X_1	X_2	X_3	...	X_{p-1}
0	0	0	0	...	0
1	1	0	0	...	0
2	0	1	0	...	0
3	0	0	1	...	0
...					
p-1	0	0	0	...	1

p groups in all: 0, 1, ..., p-1

Indicator Variables in R

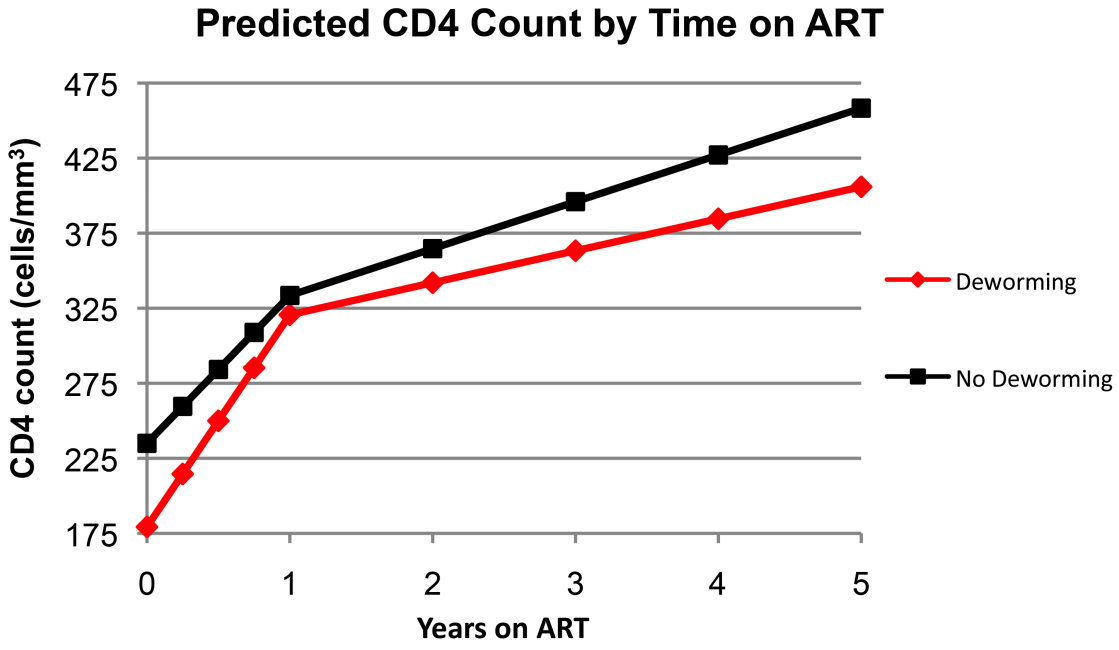
```
> str(genotype)
 num [1:30] 0 2 2 2 0 1 1 2 2 2 ...
> fitlinear<-lm(y ~ genotype)
> summary(fitlinear)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.9972     0.5829   15.44 3.19e-15 ***
genotype      5.0359     0.4157   12.12 1.19e-12 ***
---
> genotype<-factor(genotype)
> str(genotype)
 Factor w/ 3 levels "0","1","2": 1 3 3 3 1 2 2 3 3 3 ...
> fit.factor<-lm(y ~ genotype)
> summary(fit.factor)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.4246     0.3917  26.611  < 2e-16 ***
genotype1     1.7919     0.5011   3.576  0.00134 **
genotype2     9.4770     0.4929  19.226  < 2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.036 on 27 degrees of freedom
Multiple R-squared:  0.9471,Adjusted R-squared:  0.9432
F-statistic: 241.8 on 2 and 27 DF,  p-value: < 2.2e-16
```

Linear Spline ("broken arrow")

- Does the growth rate of weight for newborns slow down after 1 year?
- Does the rate of loss of CD4 cells slow down after one year of infection?
- Did the growth rate of HMOs (Health Maintenance Organizations) slow down with the enactment of new federal legislation in 1997?

This represents one way to model regression that have a "non-linear" dependence on X.

Linear Spline Example: CD4 count in HIV-infected Ugandans with antiretroviral therapy (ART)



Consider models with linear spline ("broken arrow") to account for the changes in slope at a breakpoint.

Referece: Lankowski AJ, Tsai AC, Kanyesigye M, Bwana M, Haberer JE, et al. (2014) Empiric Deworming and CD4 Count Recovery in HIV-Infected Ugandans Initiating Antiretroviral Therapy. PLoS Negl Trop Dis 8(8): e3036. doi:10.1371/journal.pntd.0003036

Defining the "spline" variable

- We define a new variable that check to see if the slope is indeed different if years on ART is greater than 1 year

$$\begin{aligned} (\text{year -1})^{+} &= (year - 1), \text{ if year on ART} > 1 \\ &= 0, \text{ if year on ART} \leq 1 \end{aligned}$$

year: Years on ART
year-1^{+}: Years since 1 year after ART treatment

	year	year1
1	0.0	0
2	0.2	0
3	0.4	0
4	0.6	0
5	0.8	0
6	1.0	0
7	1.0	0
8	2.0	1
9	3.0	2
10	4.0	3
11	5.0	4

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+$$

- What is the expected CD4 count for a patient on ART for 6 months?

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+$$

- What is the expected CD4 count for a patient on ART for 6 months?
  $\beta_0 + \beta_1 \times 0.5$
- What is the rate of CD4 increase for a patient on ART ≤ 1?

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+$$

- What is the expected CD4 count for a patient on ART for 6 months?
  $\beta_0 + \beta_1 \times 0.5$
- What is the rate of CD4 increase for a patient on ART ≤ 1?
  β_1
- What is the rate of CD4 increase for a patient on ART > 1?

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+$$

- What is the expected CD4 count for a patient on ART for 6 months?
  $\beta_0 + \beta_1 \times 0.5$
- What is the rate of CD4 increase for a patient on ART ≤ 1?
  β_1
- What is the rate of CD4 increase for a patient on ART > 1?
  $\beta_1 + \beta_2$
- What is the interpretation of β_2?

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+$$

- What is the expected CD4 count for a patient on ART for 6 months?
  $\beta_0 + \beta_1 \times 0.5$
- What is the rate of CD4 increase for a patient on ART ≤ 1?
  β_1
- What is the rate of CD4 increase for a patient on ART > 1?
  $\beta_1 + \beta_2$
- What is the interpretation of β_2?
  The difference in rates of CD4 increase between patients on ART > 1 versus ≤ 1 year (the change of slope).

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+ + ...$$

- Question: Does rate of CD increase change after one year of ART treatment in patients without deworming therapy?

Interpretation: Linear Spline

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+ + ...$$

- Question: Does rate of CD increase change after one year of ART treatment in patients without deworming therapy?
- The question in model terms is: is $\beta_2 > 0$
- Answer?

Table 2. Primary analysis: multivariable linear regression model of predictors of CD4 count (n = 5379).

Parameter	β	95% CI	p-value
Time on ART			
0 to 1 year (per year of ART up to 1 year)	98.5	85.5 to 111.6	<0.001
>1 year (per year of ART after 1 year)	31.2	26.8 to 35.6	<0.001
Age (each year of age)	−0.8	−1.4 to −0.2	0.011
TB co-infection	−114.8	−153.9 to −75.8	<0.001
Deworming	−55.6	−86.3 to −25.0	<0.001
Deworming×Time on ART interaction term†			
0 to 1 year on ART	42.8	−2.2 to 87.7	0.062
>1 year on ART	−9.9	−24.1 to 4.4	0.174

†Predicted difference in CD4 count between patients receiving versus not receiving deworming therapy in the past 90 days. The interaction terms were separated by duration of prior ART use as up to 1 year of therapy versus greater than 1 year of therapy.

doi:10.1371/journal.pntd.0003036.t002

Better Interpretation

$$E(CD4) = \beta_0 + \beta_1(year) + \beta_2(year - 1)^+ + ...$$

Table 2. Primary analysis: multivariable linear regression model of predictors of CD4 count (n = 5379).

Parameter	β	95% CI	p-value
Time on ART			
0 to 1 year (per year of ART up to 1 year)	98.5	85.5 to 111.6	<0.001
>1 year (per year of ART after 1 year)	31.2	26.8 to 35.6	<0.001
Age (each year of age)	−0.8	−1.4 to −0.2	0.011
TB co-infection	−114.8	−153.9 to −75.8	<0.001
Deworming	−55.6	−86.3 to −25.0	<0.001
Deworming×Time on ART interaction term†			
0 to 1 year on ART	42.8	−2.2 to 87.7	0.062
>1 year on ART	−9.9	−24.1 to 4.4	0.174

†Predicted difference in CD4 count between patients receiving versus not receiving deworming therapy in the past 90 days. The interaction terms were separated by duration of prior ART use as up to 1 year of therapy versus greater than 1 year of therapy.
doi:10.1371/journal.pntd.0003036.t002

- The average CD4 count for patients on ART for 1 year without deworming therapy is $\underline{\beta_0 + \beta_1}$ (95% CI: ___, __).
- For each additional year on ART, CD4 count increase 98.5 per mm^3 (95% CI: 85.5, 111.6) on average in patients on ART under 1 year.
- For each additional year on ART, CD4 count increase 31.2 per mm^3 (95% CI: 26.8, 35.6) on average in patients on ART over 1 year.

Univariate Analysis Example: Patient Characteristics

Table 1. Subject characteristics.

Characteristic	n	Received deworming at least once during study period (n = 2781)	Never received deworming during study period (n = 2598)	χ2 (p-value)
Gender; n (%)	5379			
Female	3302	1695 (61.0%)	1607 (61.9%)	0.47 (0.50)
Male	2077	1086 (39.0%)	991 (38.1%)	
Time of ART initiation; n (%)	5379			
Prior to January 1, 2007	2868	1482 (53.3%)	1386 (53.4%)	0.002 (0.97)
On or after January 1, 2007	2511	1299 (46.7%)	1212 (46.6%)	
Baseline CD4 count at time of ART initiation; median (IQR)	5379	265 (163–392)	273 (166–395)	1.72 (0.19)
Age at time of first post-ART CD4 count (years); median (IQR)	5359	38.4 (32.5–44.3)	38.0 (32.4–44.7)	1.42 (0.23)
Clinic visits at which CD4 count was obtained; median (IQR)	5379	4 (2–6)	3 (2–5)	90.49 (<0.001)
Education; n (%)	2106			
Primary only	1457	847 (69.9%)	610 (68.2%)	0.77 (0.38)
Secondary or greater	649	364 (30.1%)	285 (31.8%)	
Monthly income (Uganda shillings)*; n (%)	1439			
<100000	1095	616 (76.1%)	479 (76.0%)	0.002 (0.96)
≥100000	344	193 (23.9%)	151 (24.0%)	
Self-reported travel time from home to clinic	1577			
<1 hour	815	464 (52.6%)	351 (50.5%)	0.69 (0.41)
>1 hour	762	418 (47.4%)	344 (49.5%)	
Diagnosed with TB at least once during study period	5379			
Yes	1033	540 (19.4%)	493 (90.0%)	0.17 (0.68)
No	4346	2241 (80.6%)	2105 (81.0%)	
Pregnant at least once during study period	3302			
Yes	733	375 (22.1%)	358 (22.3%)	0.01 (0.92)
No	2569	1320 (77.9%)	1249 (77.7%)	

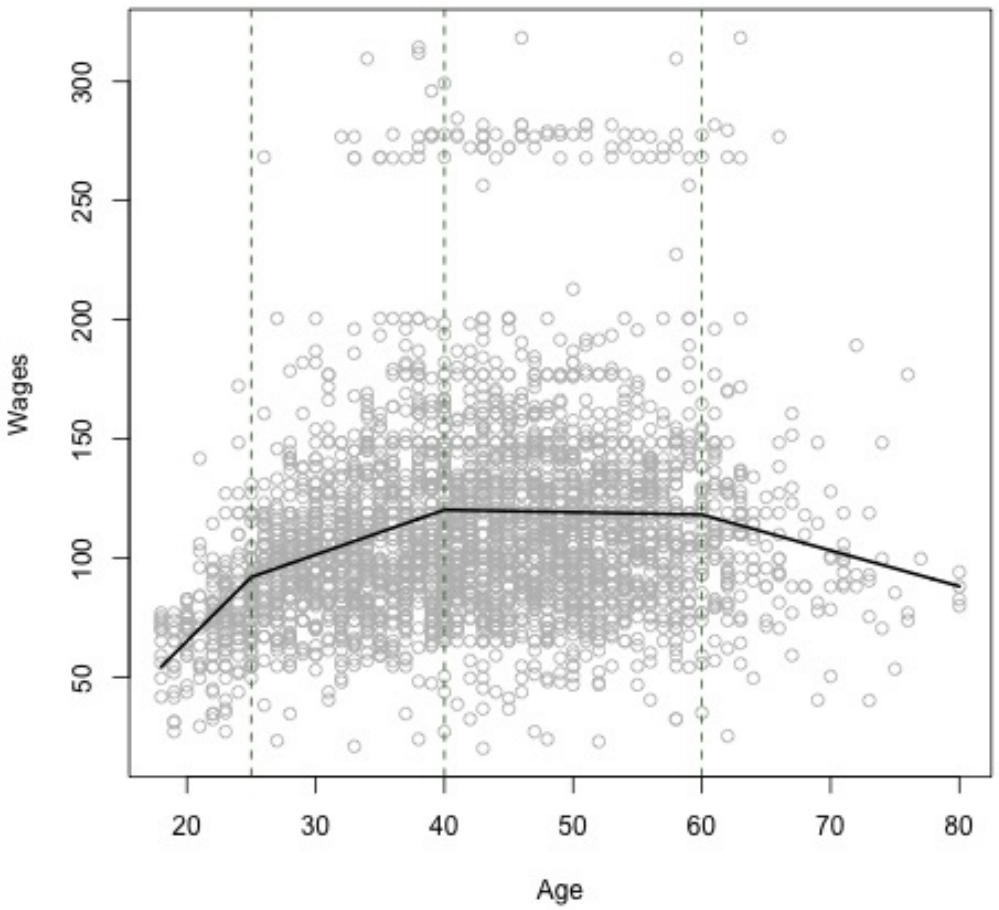
*100000 Uganda shillings valued at approximately 40 USD as of April 1, 2014.

doi:10.1371/journal.pntd.0003036.t001

Wage Example: Linear Spline

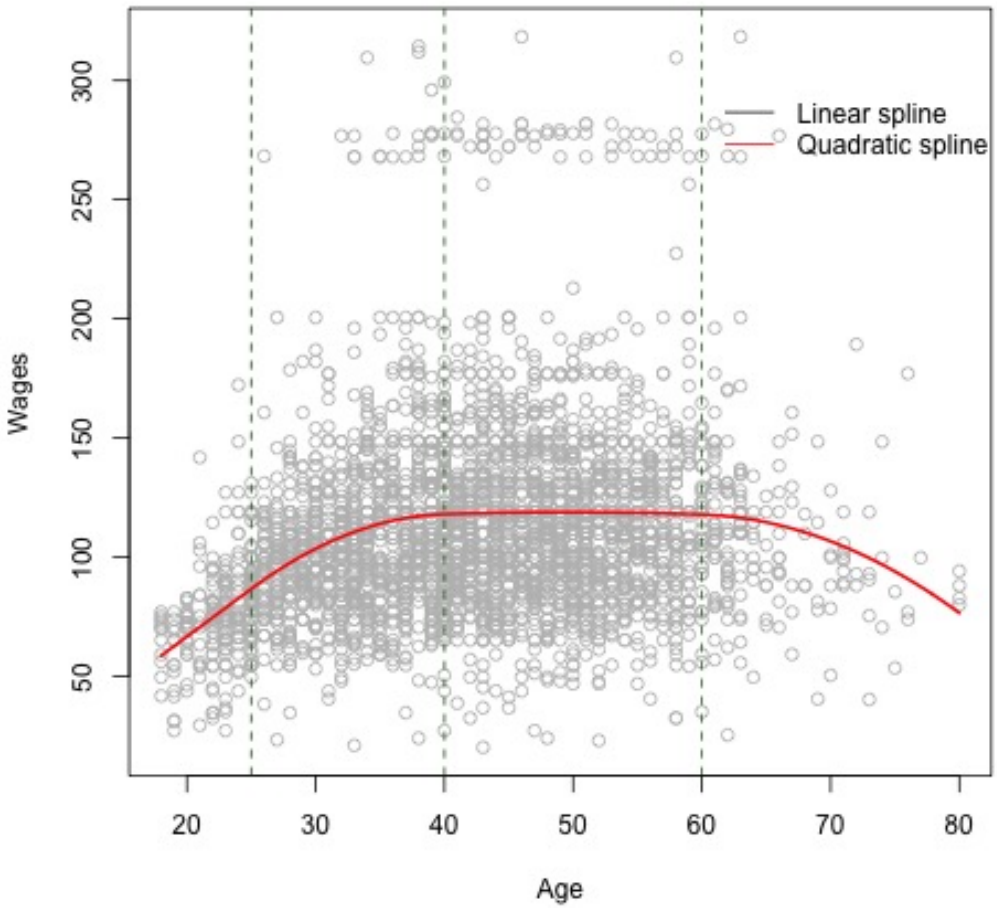
Wage and other data for a group of 3000 male workers in the Mid-Atlantic region.

$$E(wage) = \beta_0+\beta_1(age)+\beta_2(age-25)^+ +\beta_3(age-40)^+ +\beta_4(age-60)^+$$



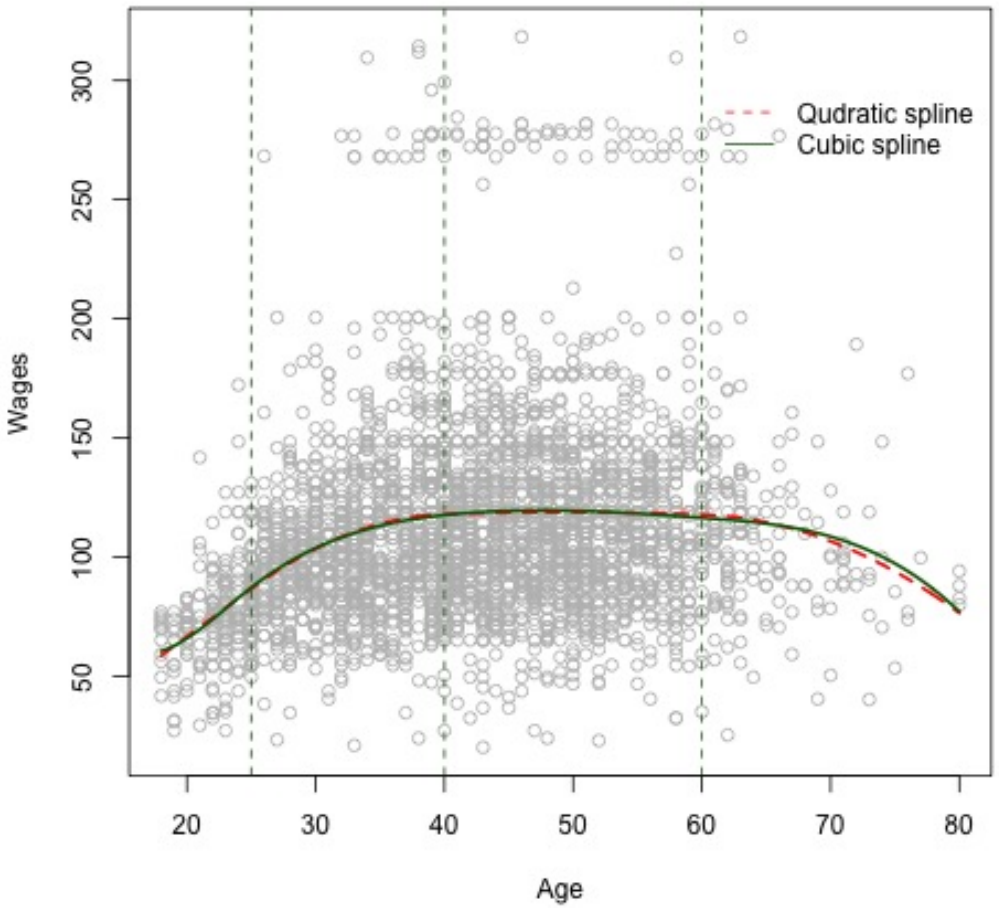
Quadratic Spline

$$\begin{aligned} E(wage) &= \beta_0 + \beta_1(age) + \beta_2(age)^2 \\ &+ \beta_3[(age-25)^+]^2 + \beta_4[(age-40)^+]^2 + \beta_5[(age-60)^+]^2 \end{aligned}$$

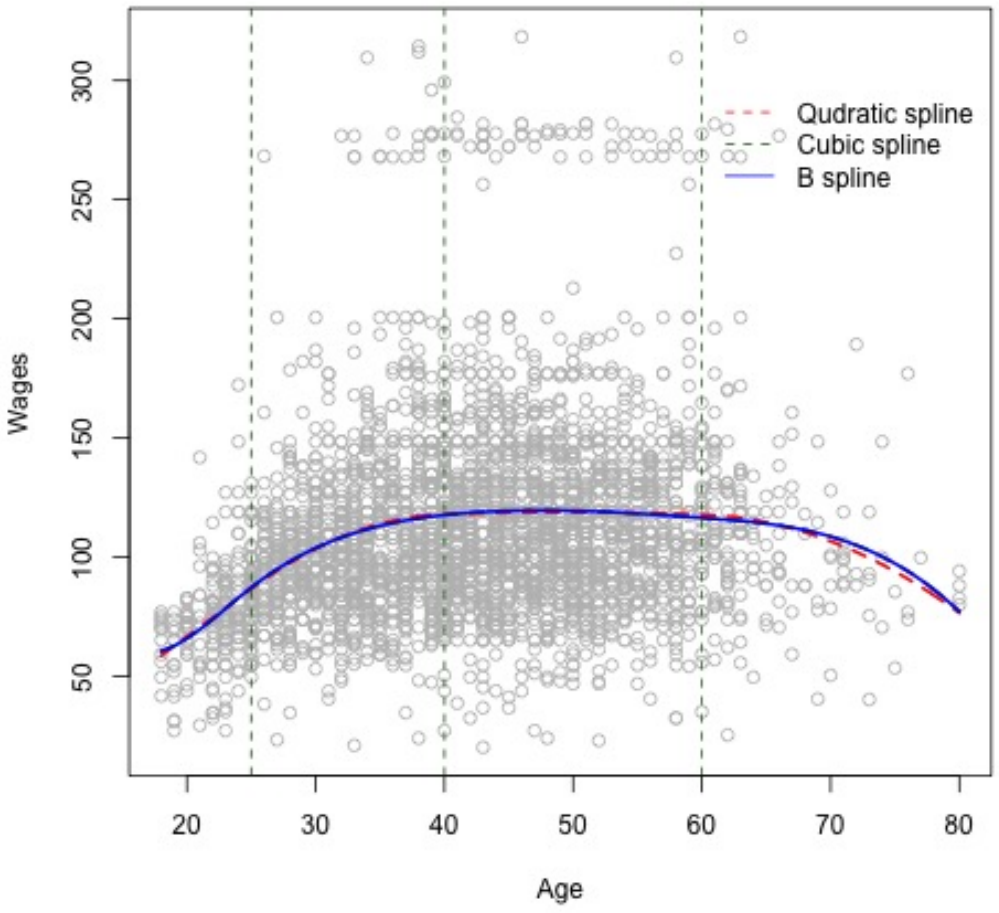


Cubic Spline

$$
\begin{aligned}
E(wage) &= \beta_0 + \beta_1(age) + \beta_2(age)^2 + \beta_3(age)^3 \\
&+ \beta_4[(age-25)^+]^3 + \beta_5[(age-40)^+]^3 + \beta_6[(age-60)^+]^3
\end{aligned}
$$



6th-order Spline



Model Selection: Akaike information criterion (AIC)

$$\text{AIC} = n\log(SSResid) - n\log(n) + 2p$$

Model Selection: Akaike information criterion (AIC)

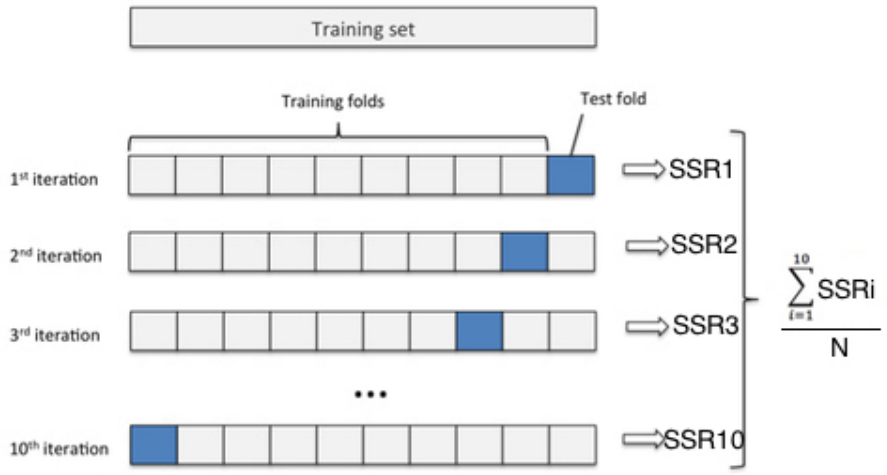
$$\text{AIC} = n\log(SSResid) - n\log(n) + 2p$$

The lower the better!

```
> AIC(fitlsp)
[1] 30639.85
> AIC(fitqsp)
[1] 30640.36
> AIC(fitcsp)
[1] 30644.59
> AIC(fitbp)
[1] 30644.59
```

Model Selection: Cross-validation

- Partition the data into random subsets.
- Leaving out one partition of the data (testing set) and estimate the parameters using the rest (training set).
- Use the fitted models to predict the Y for the left-out partition.
- Repeat this process, until all partition have fitted values.
- Calculate residual mean square using all data points.



Summary

- Interaction
  - interaction=var1 $\times$ var2
  - with interaction, the effect of one variable changes according to the level of the second variable
  - is also called "effect modification".
- Indicator (dummy) variables
  - often used to represent a categorial variable with more than 2 levels.
  - R will create dummy variables for you in *lm* if input as "factor".
- Splines are used to allow the regression line to bend.
  - often time the breakpoint is arbitrary and decided graphically
  - the actual slope above and below the breakpoint usually of more interest than the coefficient for the spline (i.e., the change in slope).
  - Increase the degree of polynomial will create smoother curves. Use AIC or cross-validation or other measures to choose good candidate models.