Notation, Notation, Notation

For all models: $s = S = \sqrt{\frac{\mathrm{SSE}}{\mathrm{df}}}$ estimate of common variance of all populations.

For a list of numbers, or a sample $X_1, \ldots, X_n$:
$SD(X) = s_x = S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

$SSX = S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2$.

For any estimator $\hat{\theta}$ for a parameter θ e.g.
$\hat{\mu}, b_0 = \hat{\beta}_0, b_1 = \hat{\beta}_1 ... \hat{Y}_i = \hat{\mu}(x_i), \hat{Y}(x^*) = \hat{\mu}(x^*)$
$\mathrm{SE}(\hat{\theta}) = s_{\hat{\theta}} = S \cdot$ something.
For example: $\mathrm{SE}(\hat{\mu}) = S\frac{1}{\sqrt{n}}, \quad \mathrm{SE}(\hat{b}_1) = S\frac{1}{\sqrt{\mathrm{SSX}}}$.

In regression: $\mathrm{SST} = \mathrm{SSY} = S_{yy}$.
$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2. \quad \mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.
$S = \sqrt{\frac{\mathrm{SSE}}{\mathrm{df}}}$ called *residual standard error* in R.

Effect of additional predictor I

Let's look at the of expenditure on sat results across different states.

```
options(digits=3)
mod1=lm(total~expend,data=sat); s1=summary(mod1)
print(s1$coefficients)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1089.3      44.39   24.54 8.17e-29
## expend         -20.9       7.33   -2.85 6.41e-03
```

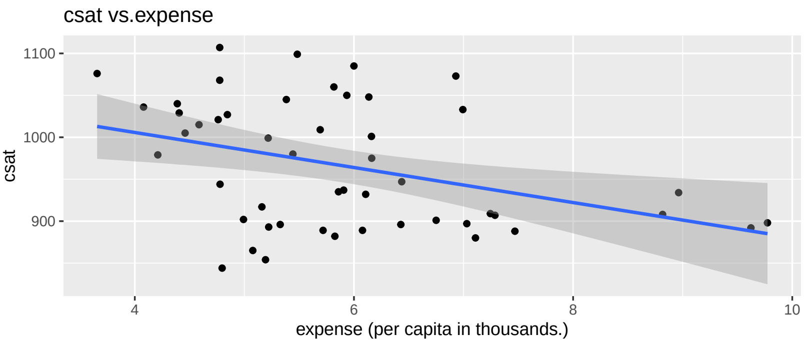
```
sprintf("sigma: %1.3f, residual df: %d, R-squared: %1.3f",s1$sigma,s1$df[2],s1$r.squared)
```

```
## [1] "sigma: 69.909, residual df: 48, R-squared: 0.145"
```

Outcome of sat's seem to decrease as a function of expenditure per student.

Effect of additional predictor II

```
qplot(expend,total,data=sat,xlab="expense (per capita in thousands.)",
      geom=c("point","smooth"),method=c("lm"),ylab="csat",
      main="csat vs.expense")
```



Effect of additional predictor III

```
mod2=lm(total~expend+takers,data=sat); s2=summary(mod2)
print(s2$coefficients)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   993.83     21.833   45.52 1.58e-40
## expend         12.29      4.224    2.91 5.53e-03
## takers         -2.85      0.215  -13.25 1.73e-17

sprintf("sigma: %1.3f, residual df: %d, R-squared: %1.3f",s2$sigma,s2$df[2],s2$r.squared)

## [1] "sigma: 32.459, residual df: 47, R-squared: 0.819"
```

Coefficient of `expend` has gone from negative to positive with inclusion of `takers`. Let's add an indicator for a region: South.

Effect of additional predictor IV

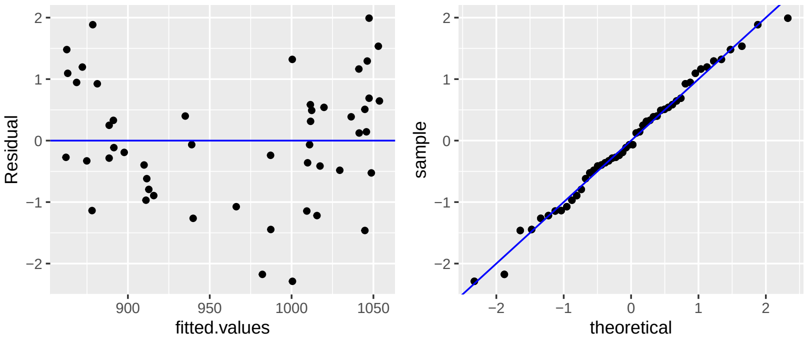
Analysis of variance between these nested models:

```
mod3=lm(total~expend+takers+I(region=="South"),data=sat)
anova(mod1,mod2,mod3)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend
## Model 2: total ~ expend + takers
## Model 3: total ~ expend + takers + I(region == "South")
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     48 234586
## 2     47  49520  1    185066 193.64 <2e-16 ***
## 3     46  43963  1      5557   5.81   0.02 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effect of additional predictor V

Let's inspect the residuals:



Multiple Binary predictors - dummy variables I

Looking at total SAT as a function of percent of takers only:

```
mod1a=lm(total~takers,data=sat); s1a=summary(mod1a)
print(s1a$coefficients)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1053.32      8.211   128.3 1.54e-62
## takers         -2.48      0.186   -13.3 9.79e-18
```

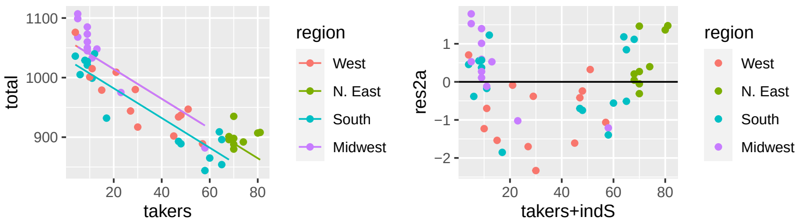
```
sprintf("sigma: %1.3f, residual df: %d, R-squared: %1.3f",s1a$sigma,s2$df[2],s1a$r.squared)
```

```
## [1] "sigma: 34.891, residual df: 47, R-squared: 0.787"
```

Multiple Binary predictors - dummy variables II

Dummy variable: indS = South/Not South

```
sat$indS=(sat$region=="South")
mod2a=lm(total~takers+indS,dat=sat)
p1=qplot(takers,total,geom=c("point"),data=sat,color=region)+
  geom_line(aes(y=predict(mod2a)))
sat$res2a=(mod2a$residuals-mean(mod2a$residuals))/sd(mod2a$residuals)
p2a=qplot(takers,res2a,data=sat,color=region,xlab="takers+indS")+geom_hline(yintercept=0)
grid.arrange(p1, p2a,  ncol=2)
```



Multiple Binary predictors - dummy variables III

Dummmy variable with 3 levels: (South, West, neither)

$x_{i2} = 1/0$ South/Not South, $x_{i3} = 1/0$ West/Not West

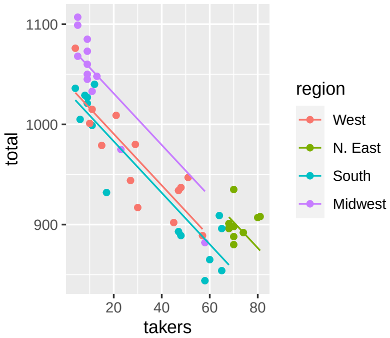
The Model:

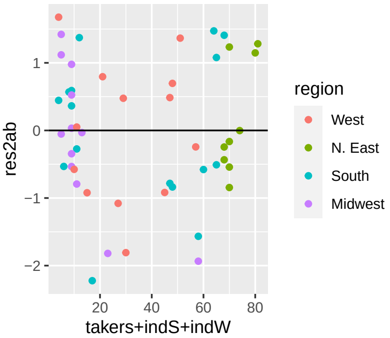
$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$

$$Y_i = \begin{cases} \beta_0 + \beta_2 + \beta_1 x_{i1} + \epsilon_i & \text{if } x_{i2} = 1, x_{i3} = 0 \\ \beta_0 + \beta_3 + \beta_1 x_{i1} + \epsilon_i & \text{if } x_{i3} = 1, x_{i2} = 0 \\ \beta_0 + \beta_1 x_{i1} + \epsilon_i & \text{if } x_{i3} = 0, x_{i2} = 0 \end{cases}$$

Multiple Binary predictors - dummy variables IV

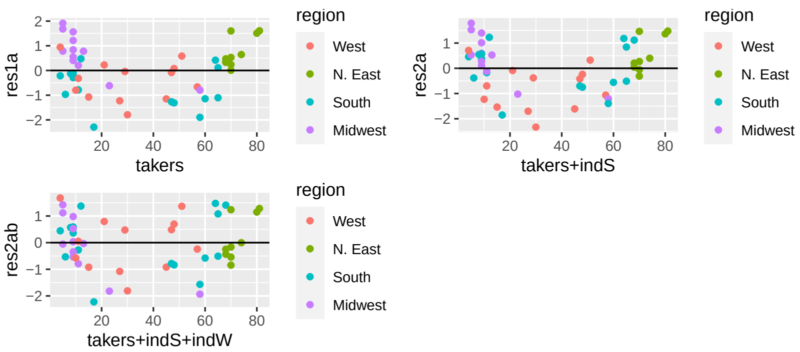
```
sat$indW=sat$region=="West"
mod2ab=lm(total~takers+indS+indW,data=sat)
sab=summary(mod2ab)
```





Multiple Binary predictors - dummy variables V

```
grid.arrange(p2,p2a, p2ab,  ncol=2)
```



Multiple Binary predictors - dummy variables VI

This isn't the same as estimating a separate regression for each region because we are assuming the *same slope* and a common variance - we are pooling data for estimating the slope and variance.

Interactions I

Add a predictor involving the product of two existing predictors:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} \cdot X_{i2} + \epsilon_i.$$

β_3 is called the interaction term.
Important: Now Y is no longer a linear function of X_1, X_2, it is a linear function of $X_1, X_2, X_1 \cdot X_2$.
When X_1 increases with X_2 fixed, Y increase depends on X_2, it is $\beta_1 + \beta_3 X_2$.

Interactions II

```
mod4=lm(total~expend+takers+expend:takers,data=sat)
summary(mod4)
```

```
##
## Call:
## lm(formula = total ~ expend + takers + expend:takers, data = sat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81.57 -25.36  -2.13  19.24  73.45
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1057.121     42.040   25.15  < 2e-16 ***
## expend           0.629      7.846    0.08    0.936
## takers          -4.232      0.818   -5.18  4.9e-06 ***
## expend:takers    0.237      0.135    1.75    0.087 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.8 on 46 degrees of freedom
## Multiple R-squared:  0.831,Adjusted R-squared:  0.82
## F-statistic: 75.2 on 3 and 46 DF,  p-value: <2e-16
```

Interactions III

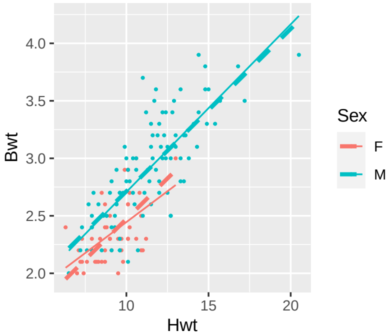
Back to cats:If one of the variables is an indicator the interaction produces a separate slope as well as having a separate intercept:
Lines go through points of means of each sub-population

```
mod=lm(Bwt~Hwt+Sex,data=cats)
modint=lm(Bwt~Hwt+Sex+Hwt:Sex,data=cats)
print(summary(modint)$coefficients)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3715     0.2734   5.016 1.57e-06
## Hwt           0.1074     0.0294   3.652 3.67e-04
## SexM         -0.1227     0.3011  -0.407 6.84e-01
## Hwt:SexM      0.0385     0.0313   1.227 2.22e-01
```

If Female: `Bwt = 1.3715 + 0.1074*Hwt`
If Male: `Bwt = 1.25 + 0.146* Hwt`

Interactions IV



This isn't the same as estimating a separate regression for each sex because we are using a common variance - we are pooling data for estimating the variance.

Polynomials I

Sometimes we want to add polynomials in the original predictors.

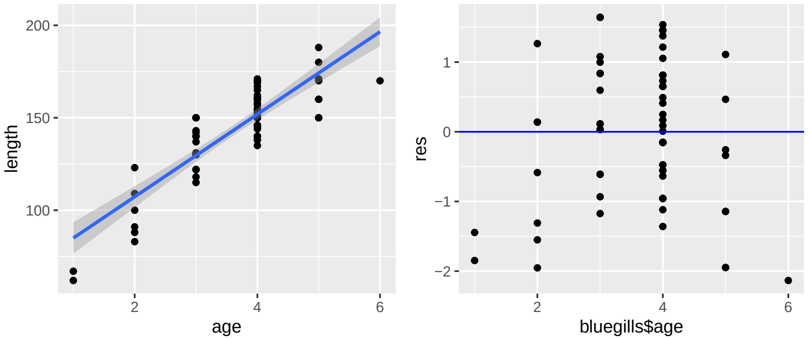
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i.$$

Again, the response is no longer linear in X, rather it is linear in X and X^2.

Polynomials II

Age and length of bluegills:

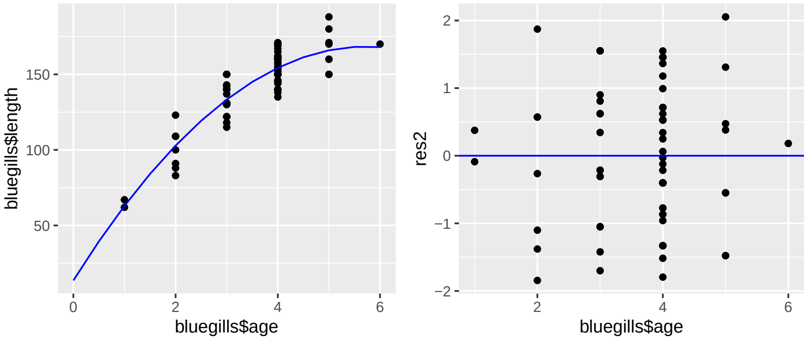
```
bluegills=read.table("bluegills.txt",header = TRUE)
mod=lm(length~age,data=bluegills)
p1=qplot(age,length,data=bluegills,geom=c("point","smooth"),method=c("lm"))
res=(mod$res-mean(mod$res))/sd(mod$res)
p2=qplot(bluegills$age,res)+geom_hline(yintercept=0,col="blue")
grid.arrange(p1,p2,ncol=2)
```



Polynomials III

Try adding a second order term:

```
mod2=lm(length~age+I(age^2),data=bluegills)
p1=qplot(bluegills$age,bluegills$length,geom=c("point"))+geom_line(aes(x=seq(0,6,.5),y=predict(mod2,newdat
res2=(mod2$res-mean(mod2$res))/sd(mod2$res)
p2=qplot(bluegills$age,res2)+geom_hline(yintercept=0,col="blue")
grid.arrange(p1,p2,ncol=2)
```



Polynomials IV

```r
summary(mod2)
```

```
##
## Call:
## lm(formula = length ~ age + I(age^2), data = bluegills)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19.85  -8.32  -1.14   6.70  22.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.622     11.016    1.24     0.22
## age           54.049      6.489    8.33  2.8e-12 ***
## I(age^2)      -4.719      0.944   -5.00  3.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.9 on 75 degrees of freedom
## Multiple R-squared:  0.801,Adjusted R-squared:  0.796
## F-statistic:  151 on 2 and 75 DF,  p-value: <2e-16
```

Polynomials V

Interactions and polynomials are exampls of **non-linear** regression.

$$Y_i = f(X_{1i}, X_{2i}) + \epsilon_i.$$

The function f is non-linear in the original variables. But it is linear in terms of polynomials defined in terms of the original variables, e.g

$$X_{1i}^2, X_{1i} \cdot X_{2i}, ...$$

.