Computing the Sums of Squares in R

```
> t(y)%*%(p2-p1)%*%y
         [,1]
[1,] 43.2
> t(y)%*%(p3-p2)%*%y
         [,1]
[1,]   42
> t(y)%*%(p4-p3)%*%y
         [,1]
[1,]  0.3
> t(y)%*%(p5-p4)%*%y
         [,1]
[1,]  2.1
> t(y)%*%(I-p5)%*%y
         [,1]
[1,] 7.48
> t(y)%*%(I-p1)%*%y
         [,1]
[1,] 95.08
```

$= y^T (P_2 - P_1) y$

$= y^T (P_3 - P_2) y$

} sizable reductions in SS

} reductions are negligible compared to SLR & quadratic regression model

end lecture 14

The ANOVA Table in R

```
> o=lm(y~x+I(x^2)+I(x^3)+I(x^4),data=d)
> anova(o)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1  43.20  43.200 57.7540 1.841e-05 ***
I(x^2)     1  42.00  42.000 56.1497 2.079e-05 ***
I(x^3)     1   0.30   0.300  0.4011    0.5407
I(x^4)     1   2.10   2.100  2.8075    0.1248
Residuals 10   7.48   0.748
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
               ' ' 1
```

matches values from Slide 41

$$\frac{MS(x|1)}{MSE} = \frac{43.20}{0.748} = 57.7540$$

What do these ANOVA F-statistics test?

1st line: Does a linear mean function fit the data significantly better than a constant mean function?

2nd line: Does a quadratic mean function fit the data significantly better than a linear mean function?

3rd line: Does a cubic mean function fit the data significantly better than a quadratic mean function?

4th line: Does a quartic mean function fit the data significantly better than a cubic mean function?

To answer each question, the error variance σ^2 is estimated from the fit of the full model with one mean for each plant density.

What do these ANOVA F-statistics test?

estimability does not necessarily imply testability

In general, we have

$$H_{0j} : (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \quad \text{vs.} \quad H_{Aj} : (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} \neq \boldsymbol{0}$$

which, in testable form, is

$$H_{0j} : \boldsymbol{C}_j\boldsymbol{\beta} = \boldsymbol{0} \quad \text{vs.} \quad H_{Aj} : \boldsymbol{C}_j\boldsymbol{\beta} \neq \boldsymbol{0},$$

where $\boldsymbol{C}_j$ is any matrix whose $q = r_{j+1} - r_j$ rows form a basis for the row space of $(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$.

in our example $q = 1$

First Line of the ANOVA Table as Test of $H_0 : C\beta = 0$

we only need 1 row to span the row space of $(P_2 - P_1)X$

linearly indep. row

```
> X=x5
> (p2-p1)%*%X
      [,1] [,2] [,3] [,4] [,5]
 [1,]  0.4  0.2    0 -0.2 -0.4
 [2,]  0.4  0.2    0 -0.2 -0.4   } * 1
 [3,]  0.4  0.2    0 -0.2 -0.4
 [4,]  0.2  0.1    0 -0.1 -0.2   } * 1/2
 [5,]  0.2  0.1    0 -0.1 -0.2
 [6,]  0.2  0.1    0 -0.1 -0.2
 [7,]  0.0  0.0    0  0.0  0.0
 [8,]  0.0  0.0    0  0.0  0.0
 [9,]  0.0  0.0    0  0.0  0.0
[10,] -0.2 -0.1    0  0.1  0.2   } * -1/2
[11,] -0.2 -0.1    0  0.1  0.2
[12,] -0.2 -0.1    0  0.1  0.2
[13,] -0.4 -0.2    0  0.2  0.4   } * -1
[14,] -0.4 -0.2    0  0.2  0.4
[15,] -0.4 -0.2    0  0.2  0.4
```

to test

$H_0 : (P_2 - P_1)X\beta = \underset{\sim}{0}$

First Line of the ANOVA Table as Test of $H_0 : \boldsymbol{C\beta} = \mathbf{0}$

Because $\text{rank}[(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{X}] = \text{rank}(\boldsymbol{P}_2 - \boldsymbol{P}_1) = \text{rank}(\boldsymbol{X}_2) - \text{rank}(\boldsymbol{X}_1) = 2 - 1 = 1$, any nonzero constant times any one nonzero row of $(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{X}$ forms a basis for the row space of $(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{X}$.

For example, we could choose $\boldsymbol{C}$ to be the following one-row matrix:

```
> 5*((p2-p1)%*%X)[15,]
[1] -2 -1  0  1  2
```

Some text books would describe these as "the coefficients of a contrast to test for linear trend." (Note this is different than a test for "lack of linear fit.")

We can add consecutive lines in an ANOVA table.

Source	Sum of Squares	DF
$x \vert 1$	$\boldsymbol{y}^\top(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$2-1=1$
$x^2 \vert 1, x$	$\boldsymbol{y}^\top(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$	$3-2=1$
$x^3 \vert 1, x, x^2$	$\boldsymbol{y}^\top(\boldsymbol{P}_4 - \boldsymbol{P}_3)\boldsymbol{y}$	$4-3=1$
$x^4 \vert 1, x, x^2, x^3$	$\boldsymbol{y}^\top(\boldsymbol{P}_5 - \boldsymbol{P}_4)\boldsymbol{y}$	$5-4=1$
Error	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_5)\boldsymbol{y}$	$15-5=10$
C. Total	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$15-1=14$

$y^T(P_5 - P_2)y$ $\quad df = 3$

We can add consecutive lines in an ANOVA table.

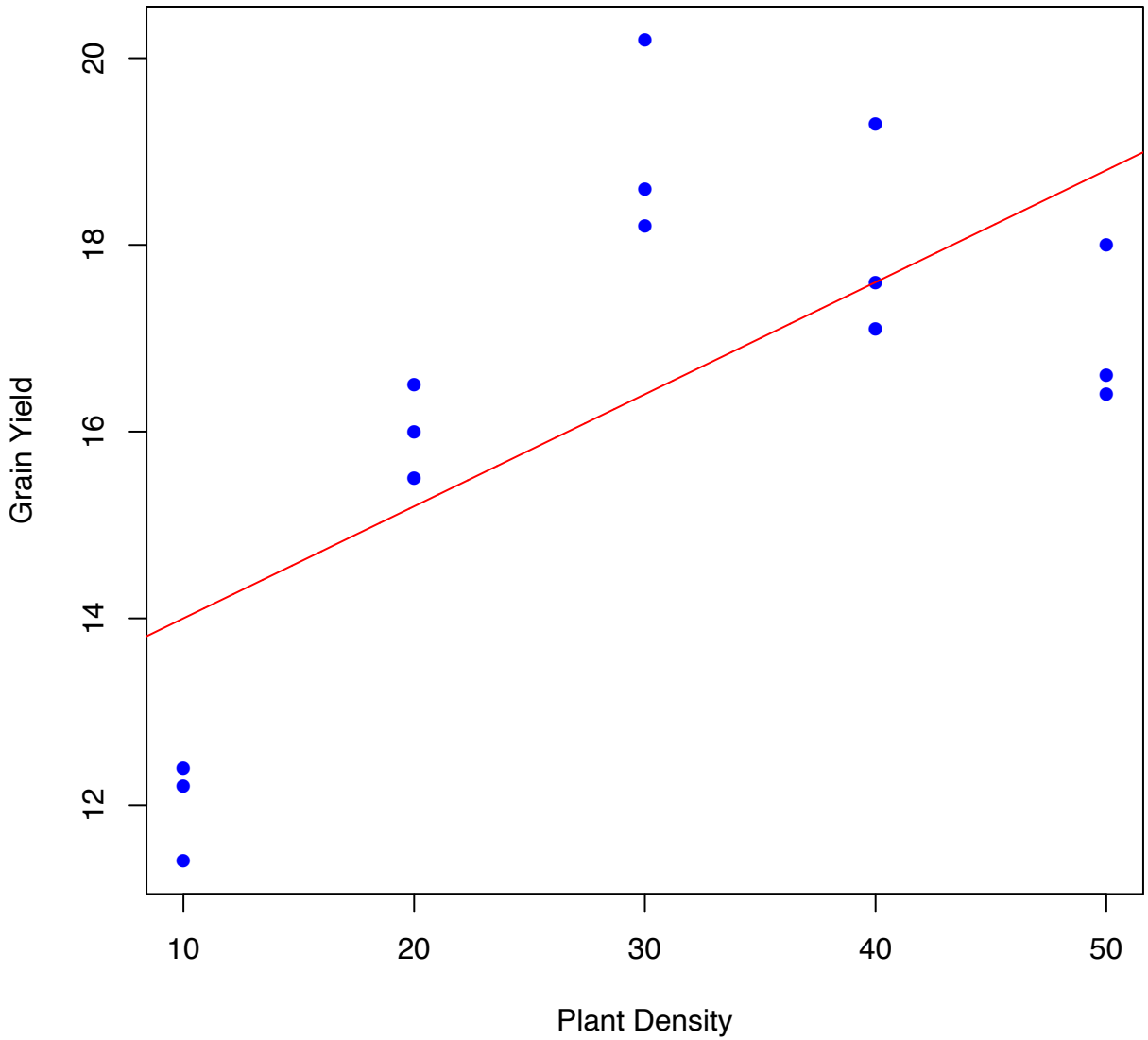
Source	Sum of Squares	DF
$x \vert 1$	$\boldsymbol{y}^\top(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$2-1=1$
$x^2, x^3, x^4, \vert 1, x$	$\boldsymbol{y}^\top(\boldsymbol{P}_5 - \boldsymbol{P}_2)\boldsymbol{y}$	$5-2=3$
Error	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_5)\boldsymbol{y}$	$15-5=10$
C. Total	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$15-1=14$

In this case, the combined rows test for lack of linear fit relative to a model with one unrestricted mean for each plant density.

Source	Sum of Squares	DF
$x \vert 1$	$\boldsymbol{y}^\top(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$2-1=1$
Lack of Linear Fit	$\boldsymbol{y}^\top(\boldsymbol{P}_5 - \boldsymbol{P}_2)\boldsymbol{y}$	$5-2=3$
Error	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_5)\boldsymbol{y}$	$15-5=10$
C. Total	$\boldsymbol{y}^\top(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$15-1=14$

we are testing whether a linear fit (SLR) is sufficient compared to a model that allows a seperate mean for each level.

```
> #Let's add the best fitting simple linear regression
> #line to our plot.
>
> o=lm(y~x,data=d)
>
> u=seq(0,60,by=.01) #overkill here but used later.
>
> lines(u,coef(o)[1]+coef(o)[2]*u,col=2)
```



$X_1 = \underline{1}$ $X_2 = [\underline{1}\ \underline{x}]$

```
> #The linear fit doesn't look very good.
> #Let's formally test for lack of fit.
>
> o=lm(y~x+factor(x),data=d)
> anova(o)
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1  43.20  43.200  57.754 1.841e-05 ***
factor(x)   3  44.40  14.800  19.786 0.0001582 ***
Residuals  10   7.48   0.748
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
                ' ' 1
```

X_5 = unrestricted means

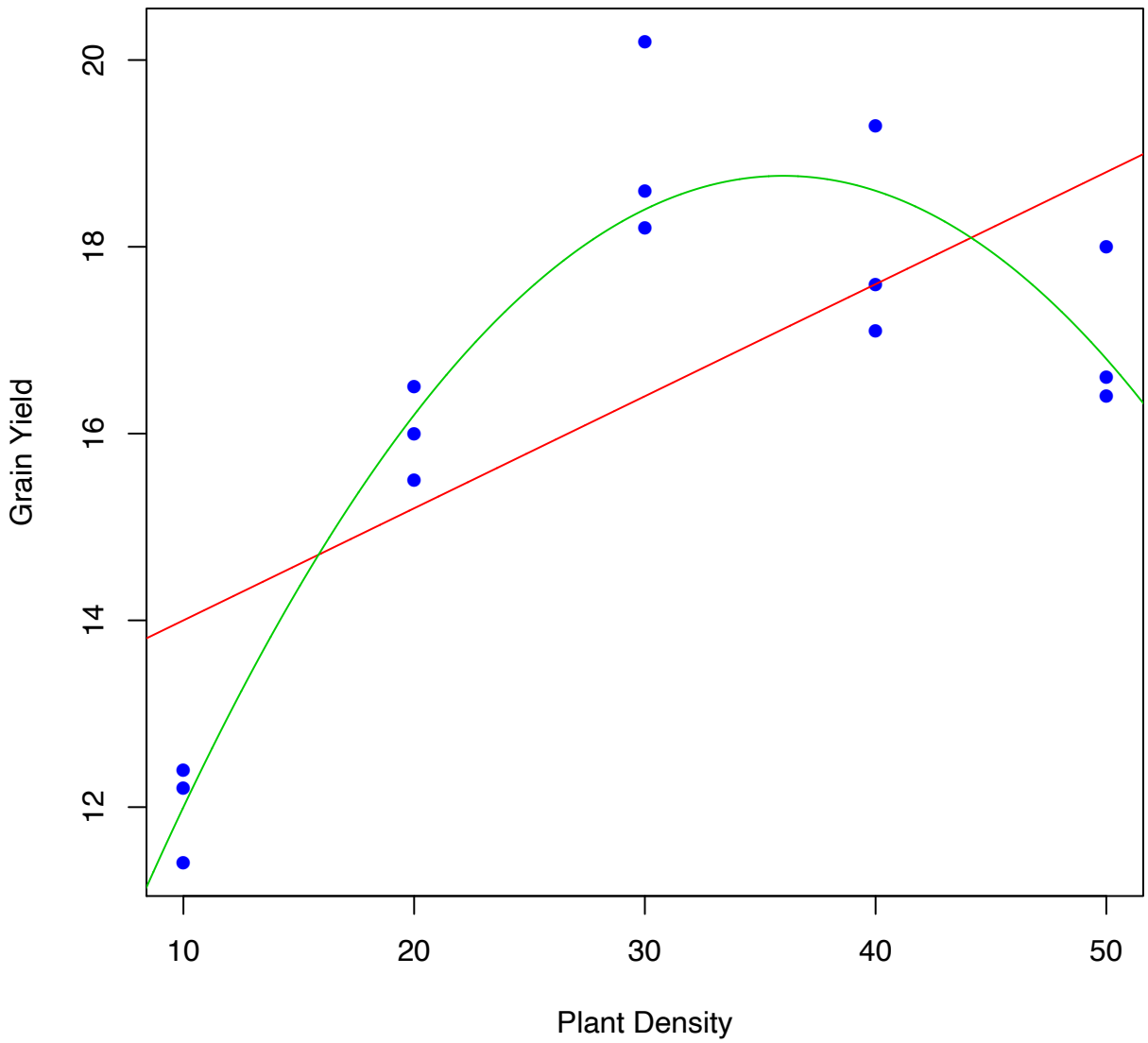
lack of linear fit suggest that the SLR is insufficient to explain all variability in the data

there is information that can be explained beyond the linear fit — we don't know what terms

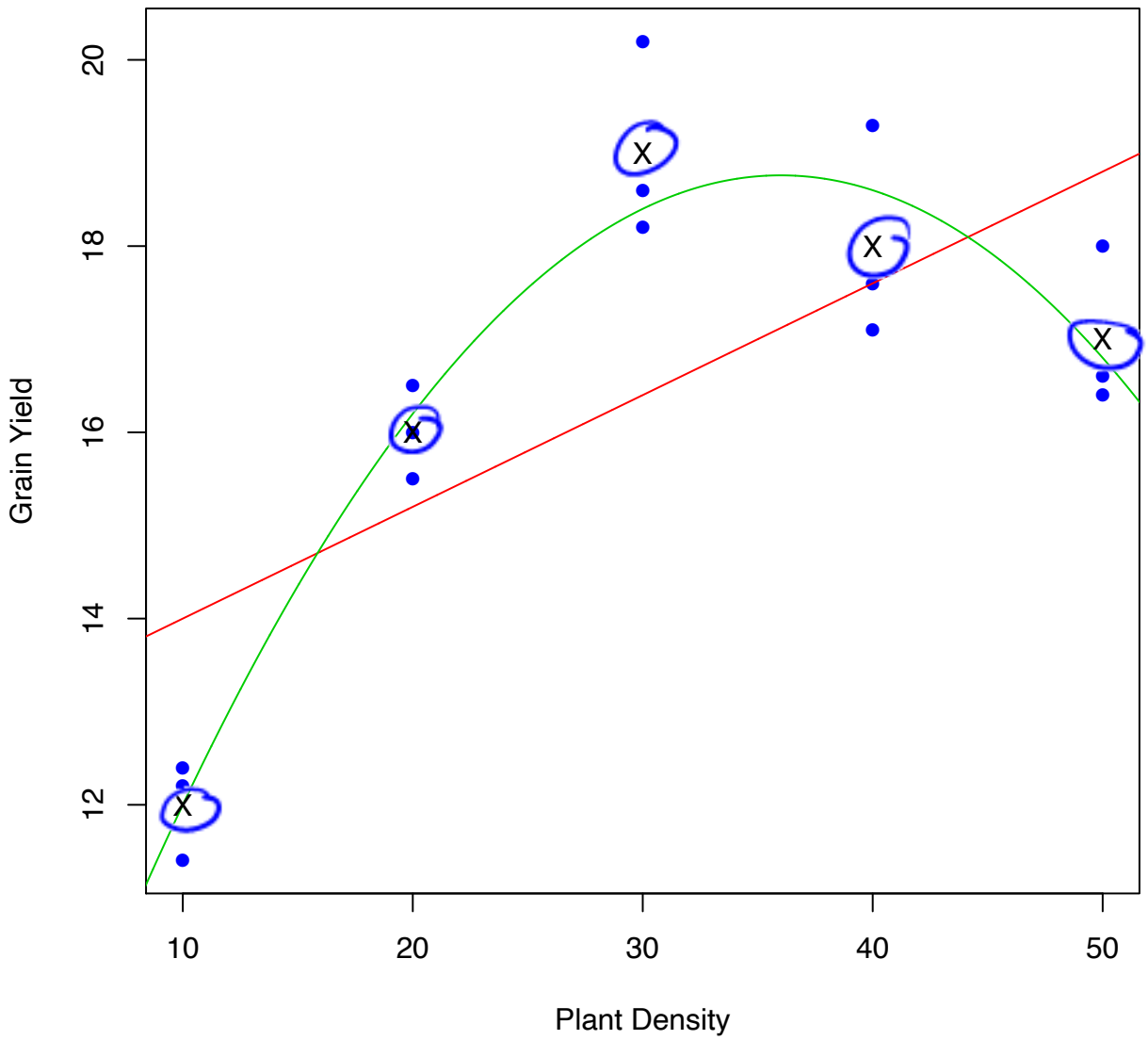
```
> #It looks like a linear fit is inadequate.
> #Let's try a quadratic fit.
>
> o=lm(y~x+I(x^2)+factor(x),data=d)
> anova(o)
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1  43.20  43.200 57.7540 1.841e-05 ***
I(x^2)      1  42.00  42.000 56.1497 2.079e-05 ***
factor(x)   2   2.40   1.200  1.6043    0.2487
Residuals  10   7.48   0.748
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
                ' ' 1
```

the quadratic term reduces the SSE significantly however the p-value of 0.2487 also suggests that a more complex is not warranted!

```
> #It looks like a quadratic fit is adequate.
> #Let's estimate the coefficients for the best
> #quadratic fit.
>
> b=coef(lm(y~x+I(x^2),data=d))
>
> #Let's add the best fitting quadratic curve
> #to our plot.
> lines(u,b[1]+b[2]*u+b[3]*u^2,col=3)
```



```
> #Let's add the treatment group means to our plot.
>
> trt.means=tapply(d$y,d$x,mean)
>
> points(unique(d$x),trt.means,pch="X")
```



```
> #The quartic fit will pass through the treatment
> #means.
>
>
> b=coef(lm(y~x+I(x^2)+I(x^3)+I(x^4),data=d))
> lines(u,b[1]+b[2]*u+b[3]*u^2+b[4]*u^3+b[5]*u^4,col=1)
```