6. ANalysis Of VAriance (ANOVA)

Setup and Notation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{I})$$

Let $\boldsymbol{X}_1 = \boldsymbol{1},\ \boldsymbol{X}_m = \boldsymbol{X},$ and $\boldsymbol{X}_{m+1} = \boldsymbol{I}$.

Suppose $\boldsymbol{X}_2,\ \ldots, \boldsymbol{X}_m$ are matrices satisfying

$$\mathcal{C}(\boldsymbol{X}_1) \subset \mathcal{C}(\boldsymbol{X}_2) \subset \cdots \subset \mathcal{C}(\boldsymbol{X}_{m-1}) \subset \mathcal{C}(\boldsymbol{X}_m).$$

Let $\boldsymbol{P}_j = \boldsymbol{P}_{\boldsymbol{X}_j}$ and $r_j = \text{rank}(\boldsymbol{X}_j)\ \forall\ j = 1, \ldots, m+1.$

The Total Sum of Squares

The *total sum of squares* (also known as the *corrected total sum of squares*) is

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y}_\cdot)^2 &= \begin{bmatrix} y_1-\bar{y}_\cdot \\ \vdots \\ y_n-\bar{y}_\cdot \end{bmatrix}' \begin{bmatrix} y_1-\bar{y}_\cdot \\ \vdots \\ y_n-\bar{y}_\cdot \end{bmatrix} = [\boldsymbol{y}-\bar{y}_\cdot\mathbf{1}]'[\boldsymbol{y}-\bar{y}_\cdot\mathbf{1}] \\
&= [\boldsymbol{y}-\boldsymbol{P}_1\boldsymbol{y}]'[\boldsymbol{y}-\boldsymbol{P}_1\boldsymbol{y}] = [\boldsymbol{I}\boldsymbol{y}-\boldsymbol{P}_1\boldsymbol{y}]'[\boldsymbol{I}\boldsymbol{y}-\boldsymbol{P}_1\boldsymbol{y}] \\
&= [(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}]'[(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}] = \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_1)'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}.
\end{aligned}$$

Partitioning the Total Sum of Squares

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y}_\cdot)^2 &= \boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{P}_{m+1}-\boldsymbol{P}_1)\boldsymbol{y} \\
&= \boldsymbol{y}'\left(\sum_{j=2}^{m+1}\boldsymbol{P}_j-\sum_{j=1}^{m}\boldsymbol{P}_j\right)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{m+1}-\boldsymbol{P}_m+\boldsymbol{P}_m-\boldsymbol{P}_{m-1}+\cdots+\boldsymbol{P}_2-\boldsymbol{P}_1)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{m+1}-\boldsymbol{P}_m)\boldsymbol{y}+\ldots+\boldsymbol{y}'(\boldsymbol{P}_2-\boldsymbol{P}_1)\boldsymbol{y} \\
&= \sum_{j=1}^{m}\boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}.
\end{aligned}
$$

The sums of squares in the equation

$$\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}$$

are often arranged in an ANOVA table.

Some Additional Sum of Squares Notation

Sum of Squares	Sum of Squares
$\boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$SS(2 \mid 1)$
$\boldsymbol{y}'(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$	$SS(3 \mid 2)$
$\vdots$	$\vdots$
$\boldsymbol{y}'(\boldsymbol{P}_m - \boldsymbol{P}_{m-1})\boldsymbol{y}$	$SS(m \mid m-1)$
$\boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_m)\boldsymbol{y}$	$SSE = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}$
$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$SSTo = \sum_{i=1}^{n}(y_i - \bar{y}_\cdot)^2$

Note that

$$
\begin{aligned}
SS(j+1 \mid j) &= \boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j + \boldsymbol{I} - \boldsymbol{I})\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_j - \boldsymbol{I} + \boldsymbol{P}_{j+1})\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_j)\boldsymbol{y} - \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{j+1})\boldsymbol{y} \\
&= SSE_j - SSE_{j+1}
\end{aligned}
$$

Thus, $SS(j+1 \mid j)$ is the amount the error sum of square decreases when $\boldsymbol{y}$ is projected onto $\mathcal{C}(\boldsymbol{X}_{j+1})$ instead of $\mathcal{C}(\boldsymbol{X}_j)$.

$SS(j+1 \mid j),\ j = 1,\ \ldots,\ m-1$ are called *Sequential Sums of Squares.*

SAS calls these *Type I Sums of Squares.*

Properties of the Matrices of the Quadratic Forms

The matrices of the quadratic forms in the ANOVA table have several useful properties:

- Symmetry
- Idempotency
- $\text{rank}(P_{j+1} - P_j) = r_{j+1} - r_j$
- Zero Cross-Products

Symmetry and Idempotency

Note that $\forall\, j = 1,\ \ldots,\ m$

$$(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)' = \boldsymbol{P}'_{j+1} - \boldsymbol{P}'_j = \boldsymbol{P}_{j+1} - \boldsymbol{P}_j$$

and

$$\begin{aligned}
(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j) &= \boldsymbol{P}_{j+1}\boldsymbol{P}_{j+1} - \boldsymbol{P}_{j+1}\boldsymbol{P}_j - \boldsymbol{P}_j\boldsymbol{P}_{j+1} + \boldsymbol{P}_j\boldsymbol{P}_j \\
&= \boldsymbol{P}_{j+1} - \boldsymbol{P}_j - \boldsymbol{P}_j + \boldsymbol{P}_j \\
&= \boldsymbol{P}_{j+1} - \boldsymbol{P}_j.
\end{aligned}$$

By idempotency and symmetry,

$$
\begin{aligned}
\boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y} &= \boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y} \\
&= [(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}]'[(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}] \\
&= ||(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}||^2 \\
&= ||\boldsymbol{P}_{j+1}\boldsymbol{y}-\boldsymbol{P}_j\boldsymbol{y}||^2 \\
&\equiv ||\hat{\boldsymbol{y}}^{(j+1)}-\hat{\boldsymbol{y}}^{(j)}||^2 \\
&= \sum_{i=1}^{n}\left(\hat{y}_i^{(j+1)}-\hat{y}_i^{(j)}\right)^2,
\end{aligned}
$$

which is why we call $\boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}$ a "sum of squares."

$\text{rank}(P_{j+1} - P_j) = r_{j+1} - r_j$

Because rank is equal to trace for idempotent matrices, we have

$$\begin{aligned}
\text{rank}(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j) &= \text{tr}(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j) = \text{tr}(\boldsymbol{P}_{j+1}) - \text{tr}(\boldsymbol{P}_j) \\
&= \text{rank}(\boldsymbol{P}_{j+1}) - \text{rank}(\boldsymbol{P}_j) \\
&= \text{rank}(\boldsymbol{X}_{j+1}) - \text{rank}(\boldsymbol{X}_j) \\
&= r_{j+1} - r_j.
\end{aligned}$$

Zero Cross-Products

$\forall\, j < \ell$

$$\begin{aligned}(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)(\boldsymbol{P}_{\ell+1} - \boldsymbol{P}_\ell) &= \boldsymbol{P}_{j+1}\boldsymbol{P}_{\ell+1} - \boldsymbol{P}_{j+1}\boldsymbol{P}_\ell - \boldsymbol{P}_j\boldsymbol{P}_{\ell+1} + \boldsymbol{P}_j\boldsymbol{P}_\ell \\ &= \boldsymbol{P}_{j+1} - \boldsymbol{P}_{j+1} - \boldsymbol{P}_j + \boldsymbol{P}_j \\ &= \mathbf{0}.\end{aligned}$$

Transposing both sides and using symmetry gives

$$(\boldsymbol{P}_{\ell+1} - \boldsymbol{P}_\ell)(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j) = \mathbf{0}.$$

Distribution of Scaled ANOVA Sums of Squares

Because

$$\left(\frac{\boldsymbol{P}_{j+1}-\boldsymbol{P}_j}{\sigma^2}\right)\left(\sigma^2\boldsymbol{I}\right)=\boldsymbol{P}_{j+1}-\boldsymbol{P}_j$$

is idempotent,

$$\frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{y}}{\sigma^2}\sim\chi^2_{r_{j+1}-r_j}\left(\frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right)$$

for all $j=1,\ \ldots,\ m$.

ANOVA Table with Degrees of Freedom

Sum of Squares	Degrees of Freedom	DF
$\boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_2) - \text{rank}(\boldsymbol{X}_1)$	$r_2 - 1$
$\boldsymbol{y}'(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_3) - \text{rank}(\boldsymbol{X}_2)$	$r_3 - r_2$
$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{y}'(\boldsymbol{P}_m - \boldsymbol{P}_{m-1})\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_m) - \text{rank}(\boldsymbol{X}_{m-1})$	$r - r_{m-1}$
$\boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_m)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_{m+1}) - \text{rank}(\boldsymbol{X}_m)$	$n - r$
$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_{m+1}) - \text{rank}(\boldsymbol{X}_1)$	$n - 1$

Mean Squares

For $j = 1, \ldots, m-1$, define

$$MS(j+1 \mid j) = \frac{SS(j+1 \mid j)}{r_{j+1} - r_j} = \frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}}{r_{j+1} - r_j}.$$

These sums of squares divided by their degrees of freedom are known as *mean squares*.

ANOVA Table with Mean Squares

Sum of Squares	Degrees of Freedom	Mean Square
$SS(2 \mid 1)$	$r_2 - 1$	$MS(2\|1)$
$SS(3 \mid 2)$	$r_3 - r_2$	$MS(3\|2)$
$\vdots$	$\vdots$	$\vdots$
$SS(m \mid m-1)$	$r - r_{m-1}$	$MS(m\|m-1)$
SSE	$n - r$	MSE
$SSTo$	$n - 1$	

Independence of ANOVA Sums of Squares

Because

$$(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\left(\sigma^2\boldsymbol{I}\right)(\boldsymbol{P}_{\ell+1} - \boldsymbol{P}_\ell) = \boldsymbol{0}$$

for all $j \neq \ell$, any two ANOVA sums of squares (not including $SSTo$) are independent.

It is also true that the ANOVA sums of squares (not including $SSTo$) are mutually independent by Cochran's Theorem, but that stronger result is not usually needed.

ANOVA F Statistics

For $j = 1, \ldots, m-1$ we have

$$F_j = \frac{MS(j+1 \mid j)}{MSE} = \frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}/(r_{j+1} - r_j)}{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}/(n-r)}$$

$$\sim F_{r_{j+1}-r_j, n-r}\left(\frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}\right).$$

ANOVA Table with F Statistics

Sum of Squares	Degrees of Freedom	Mean Square	F Stat
$SS(2 \mid 1)$	$r_2 - 1$	$MS(2\|1)$	F_1
$SS(3 \mid 2)$	$r_3 - r_2$	$MS(3\|2)$	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$SS(m \mid m-1)$	$r - r_{m-1}$	$MS(m\|m-1)$	F_{m-1}
SSE	$n - r$	MSE	
$SSTo$	$n - 1$		

Relationship with Reduced vs. Full Model F Statistic

The ANOVA F_j statistic:

$$F_j = \frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}/(r_{j+1} - r_j)}{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}/(n-r)} = \frac{MS(j+1 \mid j)}{MSE}$$

The reduced vs. full model F statistic:

$$F = \frac{\boldsymbol{y}'(\boldsymbol{P}_X - \boldsymbol{P}_{X_0})\boldsymbol{y}/(r - r_0)}{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}/(n-r)}$$

What do ANOVA F statistics test?

In general, an F statistic is used to test

H_0 : "The non-centrality parameter of the F statistic is zero."

vs.

H_A : "The non-centrality parameter of the F statistic is not zero."

What do ANOVA F statistics test?

The ANOVA F statistic

$$F_j = \frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}/(r_{j+1} - r_j)}{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}/(n - r)} = \frac{MS(j+1 \mid j)}{MSE}$$

has non-centrality parameter

$$\frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2}.$$

Thus, F_j can be used to test

$$H_{0j} : \frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2} = 0 \text{ vs. } H_{Aj} : \frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}}{2\sigma^2} \neq 0.$$

What do ANOVA F statistics test?

The following are equivalent ways to write the null and alternative hypotheses tested by F_j.

H_{0j}	H_{Aj}
$\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}=0$	$\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}\neq 0$
$(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}=\boldsymbol{0}$	$(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}\neq\boldsymbol{0}$
$\boldsymbol{P}_j E(\boldsymbol{y})=\boldsymbol{P}_{j+1}E(\boldsymbol{y})$	$\boldsymbol{P}_j E(\boldsymbol{y})\neq\boldsymbol{P}_{j+1}E(\boldsymbol{y})$
$\boldsymbol{P}_{j+1}E(\boldsymbol{y})\in\mathcal{C}(\boldsymbol{X}_j)$	$\boldsymbol{P}_{j+1}E(\boldsymbol{y})\in\mathcal{C}(\boldsymbol{X}_{j+1})\setminus\mathcal{C}(\boldsymbol{X}_j)$

What do ANOVA F statistics test?

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

is of the form

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

where $\boldsymbol{C}_j^* = (\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$.

As written, H_{0j} is not a testable hypothesis because $\boldsymbol{C}_j^*$ has n rows but rank $r_{j+1} - r_j < n$ (homework problem).

We can rewrite H_{0j} as a testable hypothesis by replacing $\boldsymbol{C}_j^*$ with any matrix $\boldsymbol{C}_j$ whose $q = r_{j+1} - r_j$ rows form a basis for the row space of $\boldsymbol{C}_j^*$.

Example: Multiple Regression

$$
\begin{aligned}
\boldsymbol{X}_1 &= \mathbf{1} \\
\boldsymbol{X}_2 &= [\mathbf{1},\ \boldsymbol{x}_1] \\
\boldsymbol{X}_3 &= [\mathbf{1},\ \boldsymbol{x}_1,\ \boldsymbol{x}_2] \\
&\vdots \\
\boldsymbol{X}_m &= [\mathbf{1},\ \boldsymbol{x}_1,\ \ldots,\ \boldsymbol{x}_{m-1}]
\end{aligned}
$$

$SS(j+1 \mid j)$ is the decrease in SSE that results when the explanatory variable x_j is added to a model containing an intercept and explanatory variables $x_1, \ldots, x_{j-1}$.

Example: Polynomial Regression

$$
\begin{aligned}
\boldsymbol{X}_1 &= \mathbf{1} \\
\boldsymbol{X}_2 &= [\mathbf{1},\ \boldsymbol{x}] \\
\boldsymbol{X}_3 &= [\mathbf{1},\ \boldsymbol{x},\ \boldsymbol{x}^2] \\
&\vdots \\
\boldsymbol{X}_m &= [\mathbf{1},\ \boldsymbol{x},\ \boldsymbol{x}^2,\ \ldots,\ \boldsymbol{x}^{m-1}]
\end{aligned}
$$

$SS(j+1 \mid j)$ is the decrease in SSE that results when the explanatory variable x^j is added to a model containing an intercept and explanatory variables $x, x^2, \ldots, x^{j-1}$.

An Example in R

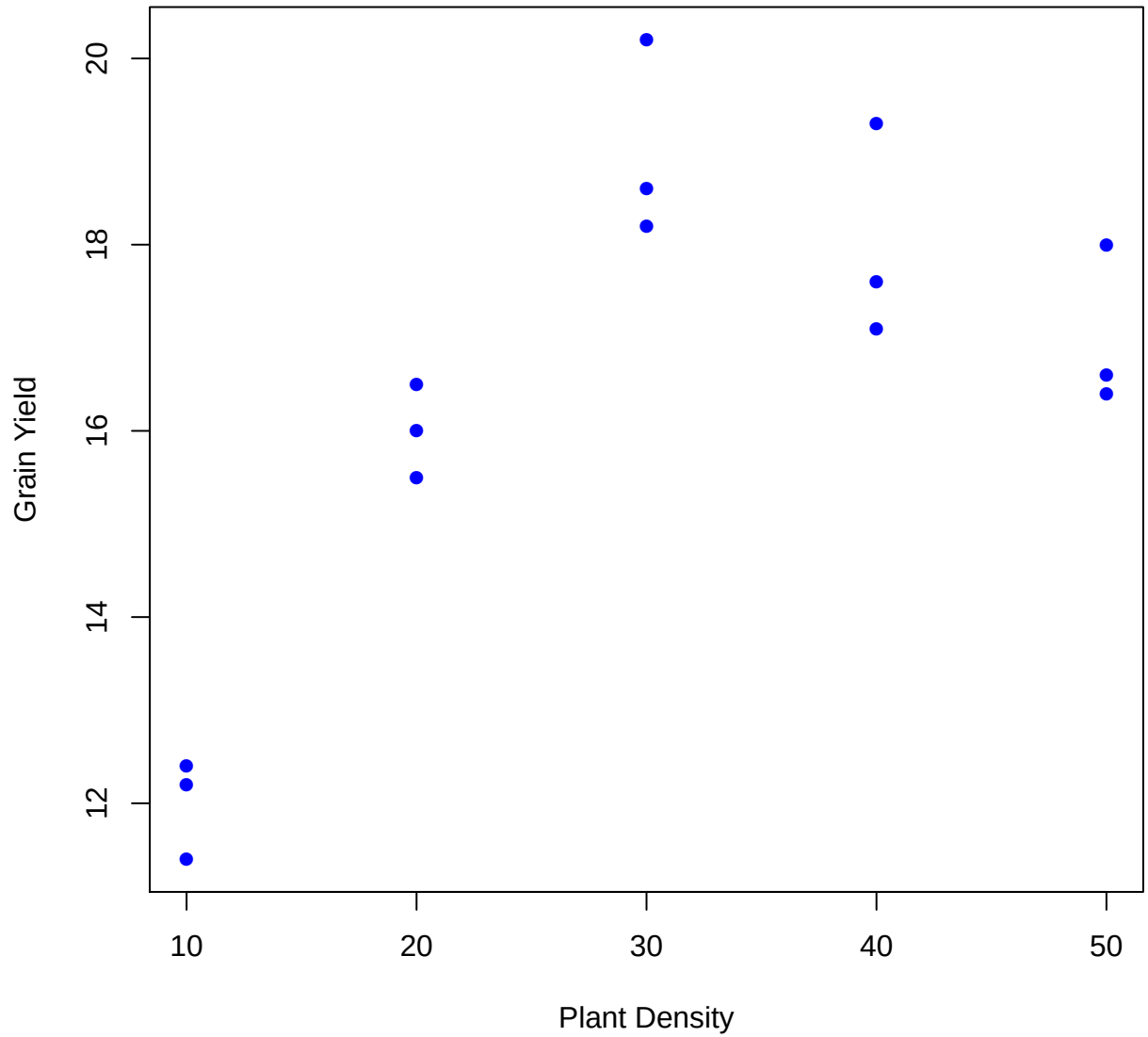
```
> #An example from "Design of Experiments: Statistical
> #Principles of Research Design and Analysis"
> #2nd Edition by Robert O. Kuehl
>
> d=read.delim("https://.../S510/PlantDensity.txt")
```

The Data

```
> d
   PlantDensity GrainYield
1            10       12.2
2            10       11.4
3            10       12.4
4            20       16.0
5            20       15.5
6            20       16.5
7            30       18.6
8            30       20.2
9            30       18.2
10           40       17.6
11           40       19.3
12           40       17.1
13           50       18.0
14           50       16.4
15           50       16.6
```

Renaming the Variables and Plotting the Data

```
> names(d)=c("x","y")
> head(d)
   x    y
1 10 12.2
2 10 11.4
3 10 12.4
4 20 16.0
5 20 15.5
6 20 16.5
>
> plot(d[,1],d[,2],col=4,pch=16,xlab="Plant Density",
+      ylab="Grain Yield")
```



Matrices with Nested Column Spaces

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} 1 & 10 \\ 1 & 10 \\ 1 & 10 \\ 1 & 20 \\ 1 & 20 \\ 1 & 20 \\ 1 & 30 \\ 1 & 30 \\ 1 & 30 \\ 1 & 40 \\ 1 & 40 \\ 1 & 40 \\ 1 & 50 \\ 1 & 50 \\ 1 & 50 \end{bmatrix}, \boldsymbol{X}_3 = \begin{bmatrix} 1 & 10 & 100 \\ 1 & 10 & 100 \\ 1 & 10 & 100 \\ 1 & 20 & 400 \\ 1 & 20 & 400 \\ 1 & 20 & 400 \\ 1 & 30 & 900 \\ 1 & 30 & 900 \\ 1 & 30 & 900 \\ 1 & 40 & 1600 \\ 1 & 40 & 1600 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 50 & 2500 \\ 1 & 50 & 2500 \end{bmatrix},$$

Matrices with Nested Column Spaces

$$\boldsymbol{X}_4 = \begin{bmatrix} 1 & 10 & 100 & 1000 \\ 1 & 10 & 100 & 1000 \\ 1 & 10 & 100 & 1000 \\ 1 & 20 & 400 & 8000 \\ 1 & 20 & 400 & 8000 \\ 1 & 20 & 400 & 8000 \\ 1 & 30 & 900 & 27000 \\ 1 & 30 & 900 & 27000 \\ 1 & 30 & 900 & 27000 \\ 1 & 40 & 1600 & 64000 \\ 1 & 40 & 1600 & 64000 \\ 1 & 40 & 1600 & 64000 \\ 1 & 50 & 2500 & 125000 \\ 1 & 50 & 2500 & 125000 \\ 1 & 50 & 2500 & 125000 \end{bmatrix}, \boldsymbol{X}_5 = \begin{bmatrix} 1 & 10 & 100 & 1000 & 10000 \\ 1 & 10 & 100 & 1000 & 10000 \\ 1 & 10 & 100 & 1000 & 10000 \\ 1 & 20 & 400 & 8000 & 160000 \\ 1 & 20 & 400 & 8000 & 160000 \\ 1 & 20 & 400 & 8000 & 160000 \\ 1 & 30 & 900 & 27000 & 810000 \\ 1 & 30 & 900 & 27000 & 810000 \\ 1 & 30 & 900 & 27000 & 810000 \\ 1 & 40 & 1600 & 64000 & 2560000 \\ 1 & 40 & 1600 & 64000 & 2560000 \\ 1 & 40 & 1600 & 64000 & 2560000 \\ 1 & 50 & 2500 & 125000 & 6250000 \\ 1 & 50 & 2500 & 125000 & 6250000 \\ 1 & 50 & 2500 & 125000 & 6250000 \end{bmatrix}$$

Centering and Standardizing for Numerical Stability

It is typically best for numerical stability to center and scale a quantitative explanatory variable prior to computing higher order terms.

In the plant density example, we could replace x by $(x - 30)/10$ and work with the matrices on the next two slides.

Because these matrices have the same column spaces as the original matrices, the ANOVA table entries are mathematically identical for either set of matrices.

Matrices with Centered and Scaled x

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & -2 \\ 1 & -2 \\ 1 & -2 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}, \boldsymbol{X}_3 = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -2 & 4 \\ 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \end{bmatrix},$$

Matrices with Centered and Scaled x

$$\boldsymbol{X}_4 = \begin{bmatrix} 1 & -2 & 4 & -8 \\ 1 & -2 & 4 & -8 \\ 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 2 & 4 & 8 \\ 1 & 2 & 4 & 8 \end{bmatrix}, \quad \boldsymbol{X}_5 = \begin{bmatrix} 1 & -2 & 4 & -8 & 16 \\ 1 & -2 & 4 & -8 & 16 \\ 1 & -2 & 4 & -8 & 16 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 & 16 \\ 1 & 2 & 4 & 8 & 16 \\ 1 & 2 & 4 & 8 & 16 \end{bmatrix}$$

Regardless of whether we center and scale x, the column space of $\boldsymbol{X}_5$ is the same as the column space of the cell means model matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

ANOVA Table for the Plant Density Data

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

Creating the Matrices in R

```
> y=d$y
> x=(d$x-mean(d$x))/10
> x
 [1] -2 -2 -2 -1 -1 -1  0  0  0  1  1  1  2  2  2
>
> n=nrow(d)
>
> x1=matrix(1,nrow=n,ncol=1)
> x2=cbind(x1,x)
> x3=cbind(x2,x^2)
> x4=cbind(x3,x^3)
> x5=matrix(model.matrix(~0+factor(x)),nrow=n)
> I=diag(rep(1,n))
```

Creating the Projection Matrices in R

```
> library(MASS)
> proj=function(x){
+   x%*%ginv(t(x)%*%x)%*%t(x)
+ }
>
> p1=proj(x1)
> p2=proj(x2)
> p3=proj(x3)
> p4=proj(x4)
> p5=proj(x5)
```

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
```

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
```

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?

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 $(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}$.

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

```
> 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
 [3,]  0.4  0.2    0 -0.2 -0.4
 [4,]  0.2  0.1    0 -0.1 -0.2
 [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
[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
[14,] -0.4 -0.2    0  0.2  0.4
[15,] -0.4 -0.2    0  0.2  0.4
```

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

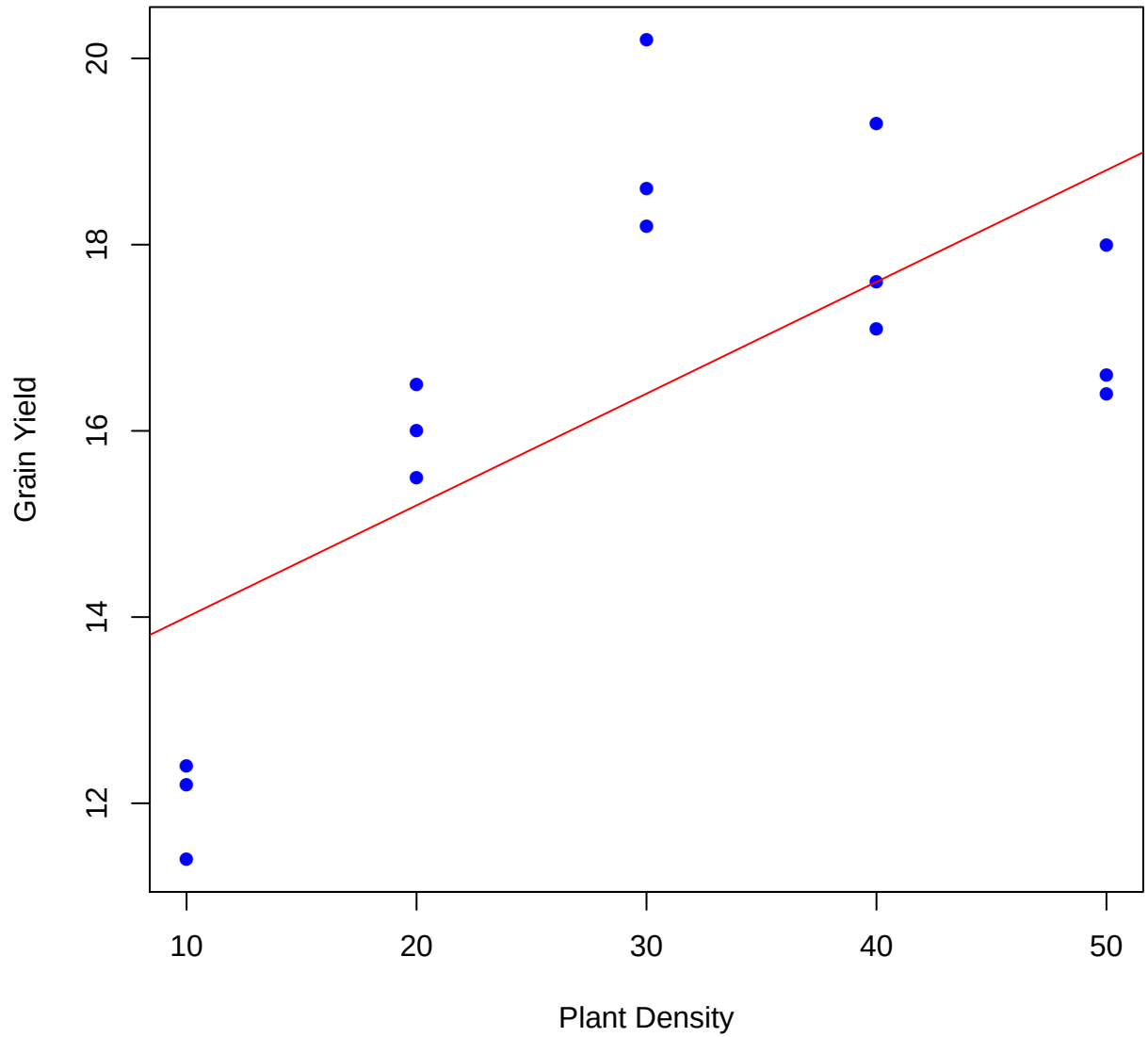
We can add consecutive lines in an ANOVA table.

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

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
$x\|1$	$\boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$2 - 1 = 1$
Lack of Linear Fit	$\boldsymbol{y}'(\boldsymbol{P}_5 - \boldsymbol{P}_2)\boldsymbol{y}$	$5 - 2 = 3$
Error	$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_5)\boldsymbol{y}$	$15 - 5 = 10$
C. Total	$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$15 - 1 = 14$

```
> #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)
```



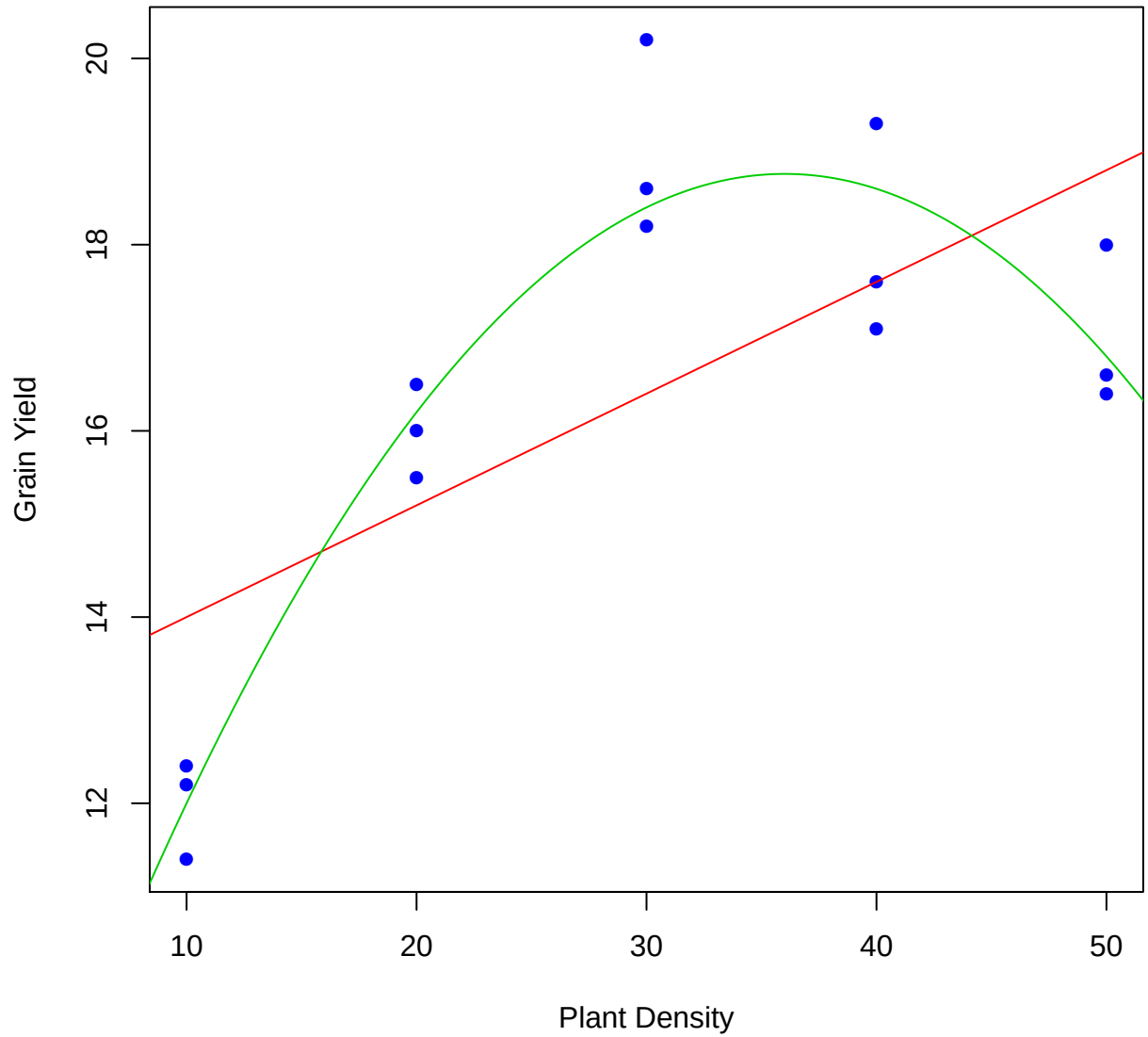
```
> #The linear fit doesn't look very good.
> #Let's formally test for lack of fit.
>
> o=lm(y~x+factor(x),data=d)
> anova(o)
Analysis of Variance Table

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
```

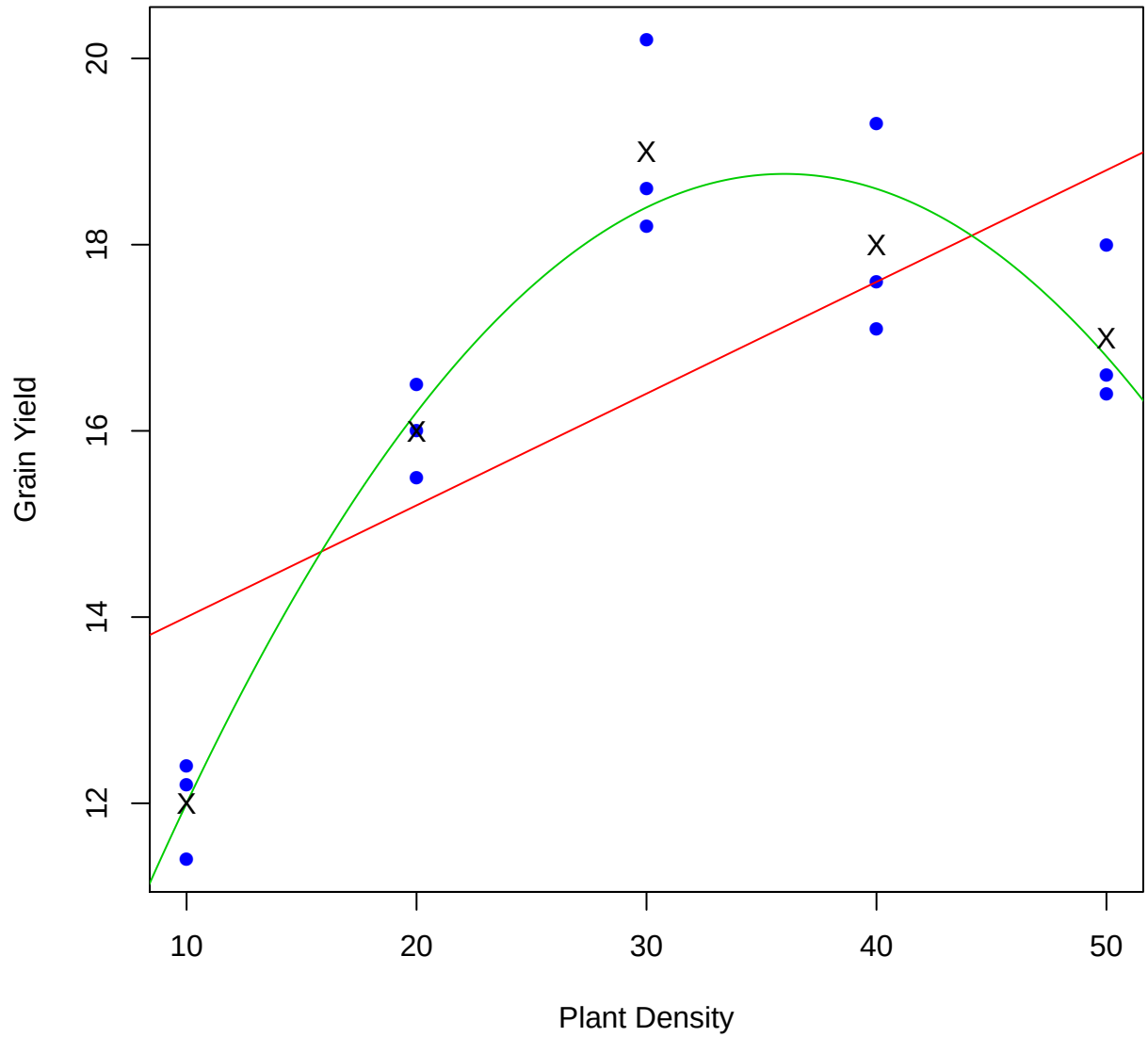
```
> #It looks like a linear fit is inadequate.
> #Let's try a quadratic fit.
>
> o=lm(y~x+I(x^2)+factor(x),data=d)
> anova(o)
Analysis of Variance Table

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
```

```
> #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)
```


Grain Yield
12
14
16
18
20
10
20
30
40
50
Plant Density

```
> #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")
```


Grain Yield
12
14
16
18
20
10
20
30
40
50
Plant Density

```
> #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)
```