Lecture 3. Experiments with a Single Factor: ANOVA

Montgomery 3-1 through 3-3

Tensile Strength Experiment

Investigate the tensile strength of a new synthetic fiber. The factor is the weight percent of cotton used in the blend of the materials for the fiber and it has five levels.

percent	tensile strength						
of cotton	1	2	3	4	5	total	average
15	7	7	11	15	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8

Data Layout for Single-Factor Experiments

treatment	observations				totals	averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a.}$	$\bar{y}_{a.}$

Analysis of Variance

- Interested in comparing several treatments.
- Could do numerous two-sample t-tests but this approach does not test equality of all treatments simultaneously.
- ANOVA provides a method of joint inference.
- Statistical Model:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2 \ldots a \\ j = 1, 2, \ldots n_i \end{cases}$$

μ - grand mean; τ_i - ith treatment effect; $\epsilon_{ij} \sim N(0, \sigma^2)$ - error

Constraint: $\sum_{i=1}^{a} n_i \tau_i = 0$.

- Basic Hypotheses:

$H_0 : \tau_1 = \tau_2 = \ldots = \tau_a = 0$ vs $H_1 : \tau_i \neq 0$ for at least one i

Derive Estimates: Partitioning y_{ij}

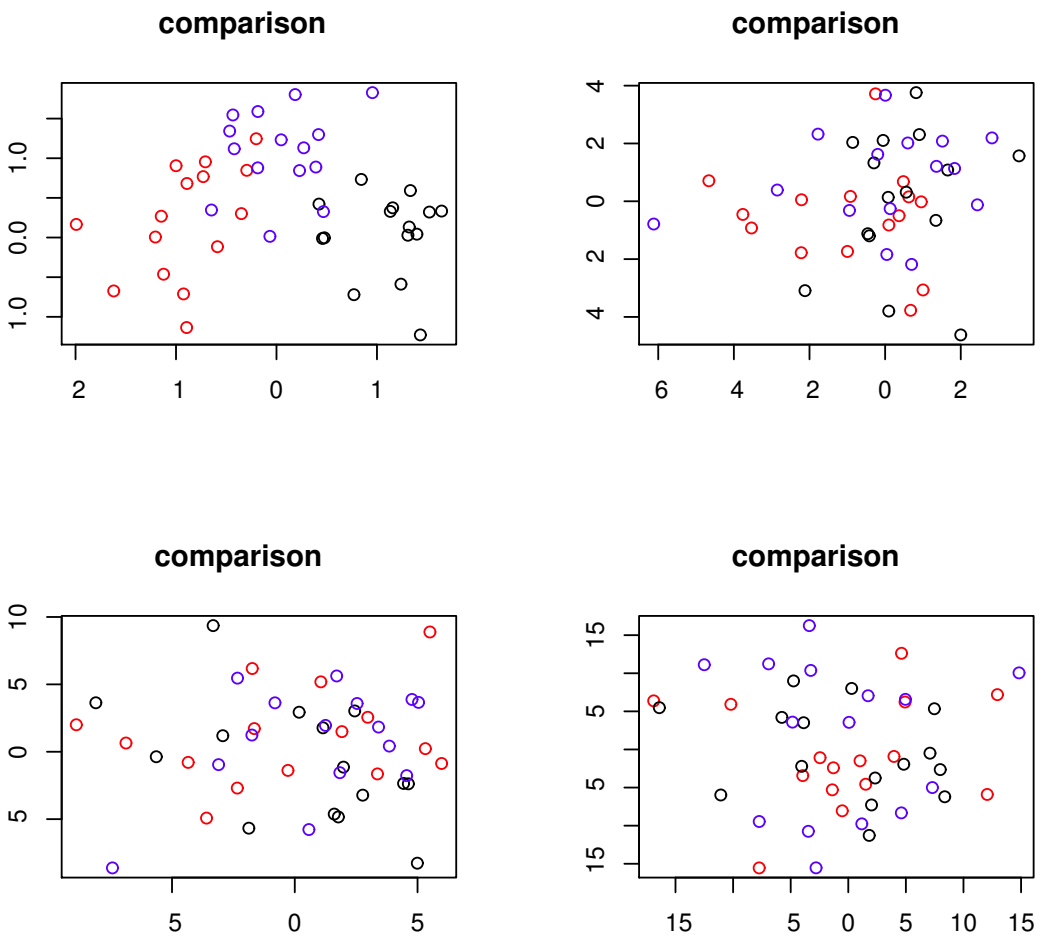
- Notation
  - $y_{i.} = \sum_{j=1}^{n_i} y_{ij} \to \overline{y}_{i.} = y_{i.}/n_i$ (treatment sample mean, or row mean)
  - $y_{..} = \sum\sum y_{ij} \to \overline{y}_{..} = y_{..}/N$ (grand sample mean)
- Decomposition of y_{ij}: $\quad y_{ij} = \overline{y}_{..} + (\overline{y}_{i.} - \overline{y}_{..}) + (y_{ij} - \overline{y}_{i.})$
- Estimates for parameters:

$$\hat{\mu} = \overline{y}_{..}$$
$$\hat{\tau}_i = (\overline{y}_{i.} - \overline{y}_{..})$$
$$\hat{\epsilon}_{ij} = y_{ij} - \overline{y}_{i.} \qquad (\text{ residual })$$

  So $y_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\epsilon}_{ij}$.
  - It can be verified that

    $\sum_{i=1}^{a} n_i \hat{\tau}_i = 0;$ $\qquad$ $\sum_{j=1}^{n_i} \hat{\epsilon}_{ij} = 0$ for all i.

The intuition



Test Basic Hypotheses: Partitioning the Sum of Squares

- Recall $y_{ij} - \overline{y}_{..} = \hat{\tau}_i + \hat{\epsilon}_{ij} = (\overline{y}_{i.} - \overline{y}_{..}) + (y_{ij} - \overline{y}_{i.})$
- Can show

$$\begin{aligned} \sum_i \sum_j (y_{ij} - \overline{y}_{..})^2 = & \quad \sum_i n_i (\overline{y}_{i.} - \overline{y}_{..})^2 \quad + \sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2 \\ = & \quad \sum_i n_i \hat{\tau}_i^2 \quad + \sum_i \sum_j \hat{\epsilon}_{ij}^2. \\ \text{Total SS} = & \quad \text{Treatment SS} \quad + \text{Error SS} \\ \text{Total Variation} = & \quad \text{Variation between} \quad + \text{Variation within} \end{aligned}$$

  $\mathrm{SS_T}$ = $\mathrm{SS_{Treatments}}$ + $\mathrm{SS_E}$

- Derive test statistics for testing $H_0 : \tau_1 = \cdots = \tau_a = 0$:
  look at $\mathrm{SS_{Treatments}}$= $\sum n_i \hat{\tau}_i^2 = \sum n_i (\overline{y}_{i.} - \overline{y}_{..})^2$

  Small if $|\hat{\tau}_i|'s$ are small

  If large then reject H_0, but how large is large?

  Standardize to account for inherent variability

Test Statistic

$$F_0 = \frac{\mathrm{SS}_{\mathrm{Treatments}}/(a-1)}{\mathrm{SS}_{\mathrm{E}}/(N-a)} = \frac{\mathrm{MS}_{\mathrm{Treatments}}}{\mathrm{MS}_{\mathrm{E}}}$$

- Mean squares:

  $\mathrm{MS}_{\mathrm{Treatments}} = \sum_i n_i \hat{\tau}_i^2/(a-1)$ $\mathrm{MS}_{\mathrm{E}} = \sum_i \sum_j \hat{\epsilon}_{ij}^2/(N-a)$

- Can show that the expected values are

  $\mathrm{E}(\mathrm{MS}_{\mathrm{E}})=\sigma^2$

  $\mathrm{E}(\mathrm{MS}_{\mathrm{Treatment}}) = \sigma^2 + \sum n_i \tau_i^2/(a-1)$

- Hence, MS_{E} is always a good (unbiased) estimator for σ^2, that is, $\hat{\sigma}^2 = \mathrm{MS}_{\mathrm{E}}$.

  Under H_0, $\mathrm{MS}_{\mathrm{Treatment}}$ is also a good (unbiased) estimator for σ^2.

  Under H_0, F_0 is expected close to 1. Large F_0 implies deviation from H_0.

- What is the sampling distribution of F_0 under H_0?

Sampling Distribution of F_0 under H_0

- Based on model:

  $y_{ij} \sim \mathrm{N}(\mu + \tau_i, \sigma^2)$ implies $\overline{y}_{i.} \sim \mathrm{N}(\mu + \tau_i, \sigma^2/n_i)$

  implies $\sum_j (y_{ij} - \overline{y}_{i.})^2/\sigma^2 \sim \chi^2_{n_i - 1}$

  Therefore:

  $\mathrm{SS_E}/\sigma^2 = \sum_i \sum_j (y_{ij} - \overline{y}_{i.})^2/\sigma^2 \sim \chi^2_{N-a}$

- Under H_0:

  $\mathrm{SS_{Treatment}}/\sigma^2 = \sum_i n_i(\overline{y}_{i.} - \overline{y}_{..})^2/\sigma^2 \sim \chi^2_{a-1}$

- $\mathrm{SS_E}$ and $\mathrm{SS_{Treatment}}$ are independent. So

$$F_0 = \frac{\mathrm{SS_{Treatment}}/\sigma^2(a-1)}{\mathrm{SS_E}/\sigma^2(N-a)} = \frac{\chi^2_{a-1}/(a-1)}{\chi^2_{N-a}/(N-a)} \sim F_{a-1,N-a}$$

  F-distribution with numerator df $a-1$ and denominator df $N-a$.

Analysis of Variance (ANOVA) Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Between	$\text{SS}_{\text{Treatment}}$	$a-1$	$\text{MS}_{\text{Treatment}}$	F_0
Within	SS_{E}	$N-a$	MS_{E}	
Total	SS_{T}	$N-1$		

If balanced:

$$\text{SS}_{\text{T}} = \sum\sum y_{ij}^2 - y_{..}^2/N; \qquad \text{SS}_{\text{Treatment}} = \frac{1}{n}\sum y_{i.}^2 - y_{..}^2/N$$

SS_{E}=SS_{T} - $\text{SS}_{\text{Treatment}}$

If unbalanced:

$$\text{SS}_{\text{T}} = \sum\sum y_{ij}^2 - y_{..}^2/N; \qquad \text{SS}_{\text{Treatment}} = \sum \frac{y_{i.}^2}{n_i} - y_{..}^2/N$$

SS_{E}=SS_{T} - $\text{SS}_{\text{Treatment}}$

- **Decision Rule**: If $F_0 > F_{\alpha,a-1,N-a}$ then reject H_0

Connection between F-Test and Two-Sample t-Test

- a=2
- Consider the square of the t-test statistic

$$t_0^2 = \frac{(\bar{y}_{1.} - \bar{y}_{2.})^2}{(s_{pool}\sqrt{1/n_1 + 1/n_2})^2} = \frac{\frac{n_1 n_2}{n_1+n_2}(\bar{y}_{1.} - \bar{y}_{2.})^2}{s_{pool}^2}$$

$$= \frac{\frac{n_1 n_2}{n_1+n_2}[(\bar{y}_{1.} - \bar{y}_{..}) - (\bar{y}_{2.} - \bar{y}_{..})]^2}{s_{pool}^2}$$

$$= \frac{\frac{n_1 n_2}{n_1+n_2}(\bar{y}_{1.} - \bar{y}_{..})^2 + \frac{n_1 n_2}{n_1+n_2}(\bar{y}_{2.} - \bar{y}_{..})^2 - \frac{2n_1 n_2}{n_1+n_2}(\bar{y}_{1.} - \bar{y}_{..})(\bar{y}_{2.} - \bar{y}_{..})}{s_{pool}^2}$$

Consider

$$\bar{y}_{..} = \frac{1}{n_1+n_2}(\sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j}) = \frac{n_1}{n_1+n_2}\bar{y}_{1.} + \frac{n_2}{n_1+n_2}\bar{y}_{2.}$$

One has

$$t_0^2 = \frac{[n_1(\bar{y}_{1.} - \bar{y}_{..})^2 + n_2(\bar{y}_{2.} - \bar{y}_{..})^2]/(2-1)}{[\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_{1.})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_{2.})^2]/(n_1 + n_2 - 2)}$$

$$= \frac{\mathrm{MS}_{\mathrm{Treatment}}}{\mathrm{MS}_{\mathrm{E}}}$$

- When $a = 2$, $t_0^2 = F_0$
- When $a = 2$, F-test and two-sample two-sided test are equivalent.

Example

Twelve lambs are randomly assigned to three different diets. The weight gain (in two weeks) is recorded. Is there a difference between the diets?

Diet 1	8	16	9		
Diet 2	9	16	21	11	18
Diet 3	15	10	17	6	

- $N = 12$, $\sum\sum y_{ij} = 156$ and $\bar{y}_{..} = 156/12 = 13$.
- $n_1 = 3, y_{1.} = 33, \bar{y}_{1.} = 11; n_2 = 5, y_{2.} = 75, \bar{y}_{2.} = 15; n_3 = 4,$ $y_{3.} = 48$ and $\bar{y}_{3.} = 12$.
- $\hat{\tau}_1 = \bar{y}_{1.} - \bar{y}_{..} = 11 - 13 = -2$; Similarly, $\hat{\tau}_2 = 15 - 13 = 2$ and $\hat{\tau}_3 = 12 - 13 = -1$.
- $\mathrm{SS_T} = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 246$.
- $\mathrm{SS_{Treatment}} = 3 * (-2)^2 + 5 * (2)^2 + 4 * (-1)^2 = 36$.
- $\mathrm{SS_E} = 246 - 36 = 210$

- $\mathrm{MS}_E = \hat{\sigma}^2 = 210/(12-3) = 23.33$
- $F_0 = (36/2)/(210/9) = 0.77$; P-value > 0.25
- Fail to reject $H_0 : \tau_1 = \tau_2 = \tau_3 = 0$.

Using SAS (lambs.sas)

```
option nocenter ps=65 ls=80;

data lambs;
 input diet wtgain;
 datalines;
  1  8
  1 16
  1  9
  2  9
  2 16
  2 21
  2 11
  2 18
  3 15
  3 10
  3 17
  3  6
;
```

```
symbol1 bwidth=5 i=box; axis1 offset=(5);
proc gplot; plot wtgain*diet / frame haxis=axis1;

proc glm;
 class diet;
 model wtgain=diet;
 output out=diag r=res p=pred;

proc gplot; plot res*diet /frame haxis=axis1;

proc sort; by pred;
symbol1 v=circle i=sm50;
proc gplot; plot res*pred / haxis=axis1;
run;
```

SAS Output

The GLM Procedure
Dependent Variable: wtgain

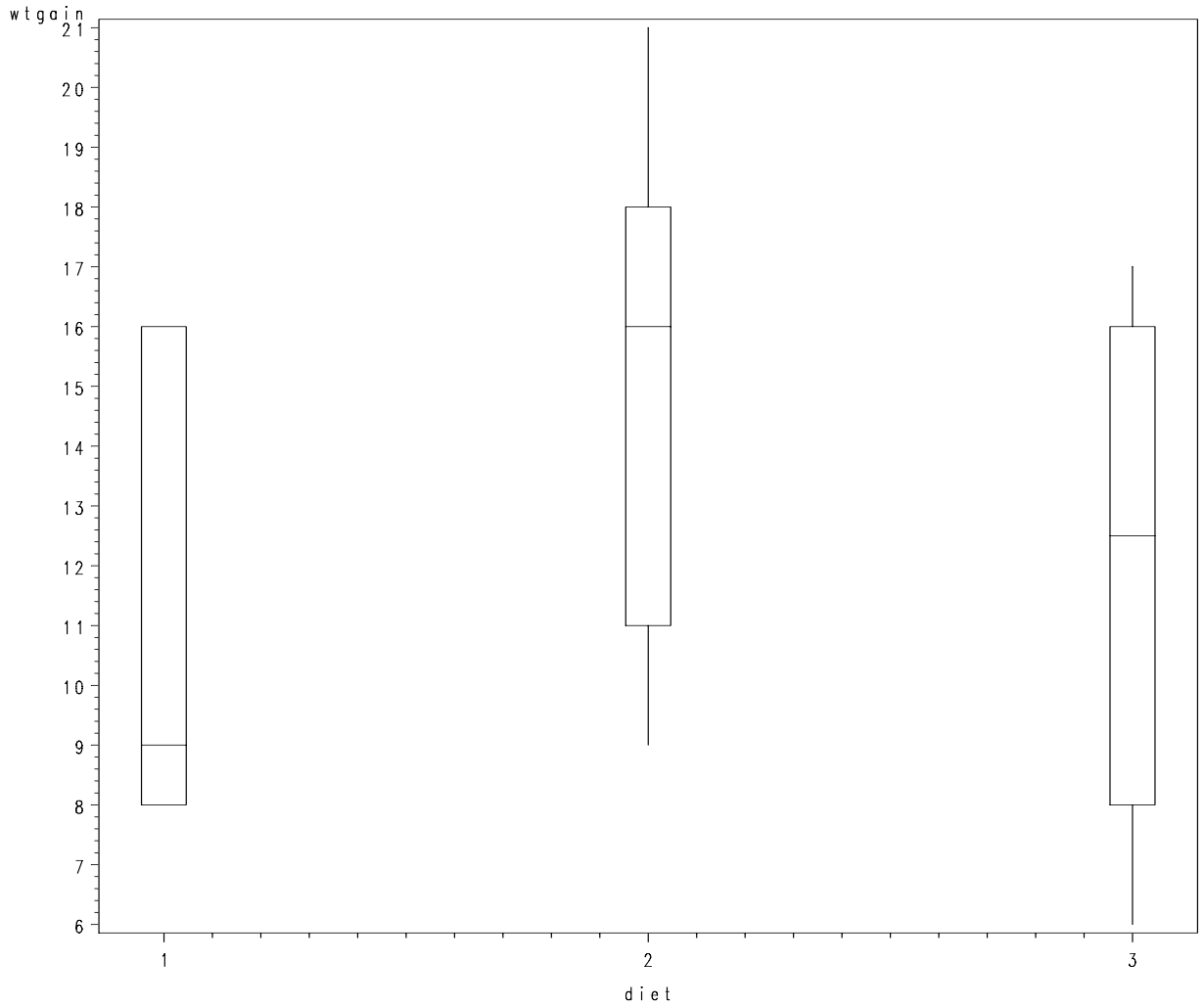
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	36.0000000	18.0000000	0.77	0.4907
Error	9	210.0000000	23.3333333		
Corrected Total	11	246.0000000			

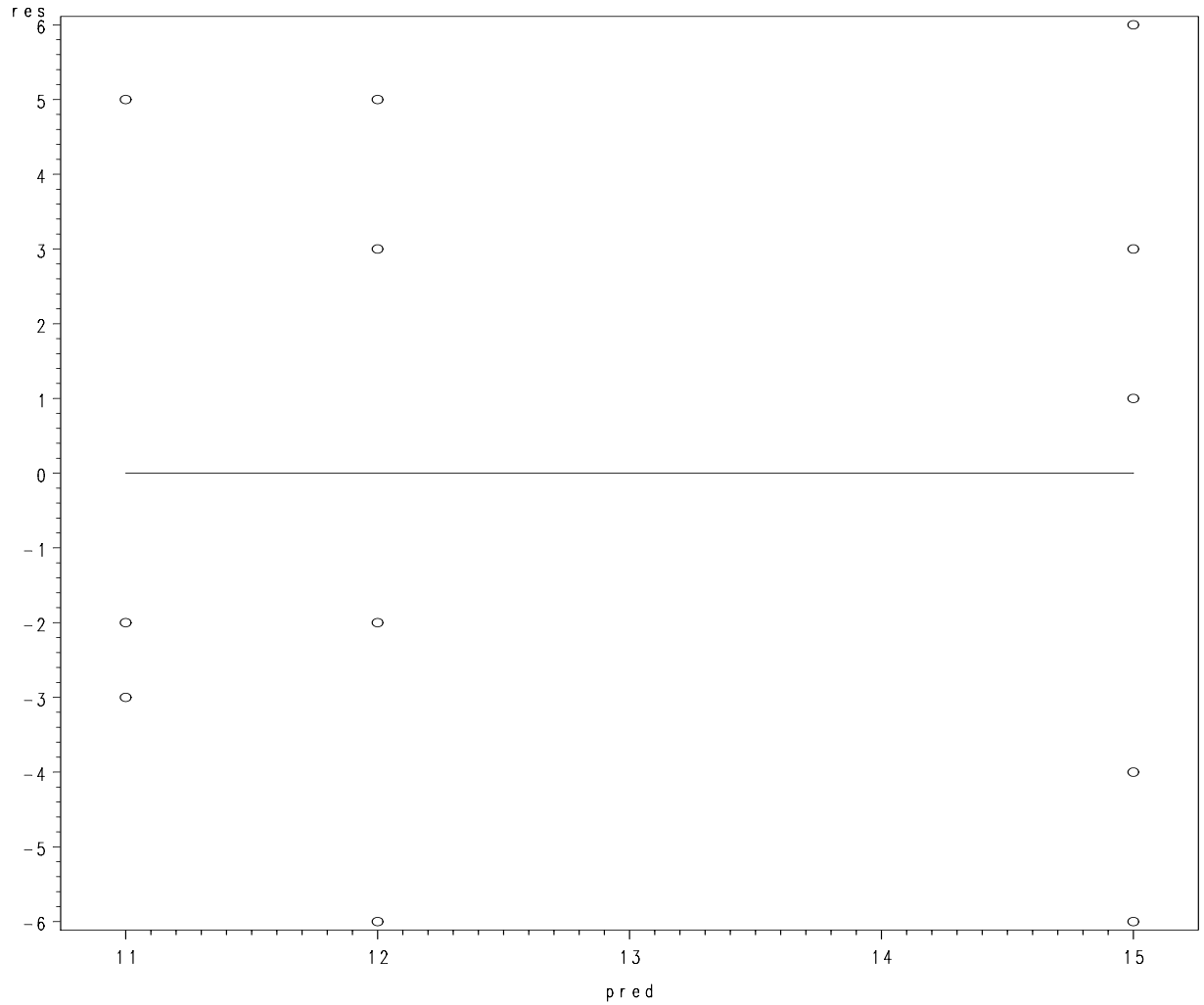
R-Square	Coeff Var	Root MSE	wtgain Mean
0.146341	37.15738	4.830459	13.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
diet	2	36.00000000	18.00000000	0.77	0.4907

Source	DF	Type III SS	Mean Square	F Value	Pr > F
diet	2	36.00000000	18.00000000	0.77	0.4907

Side-by-Side Boxplot





Residual Plot

Tensile Strength Experiment

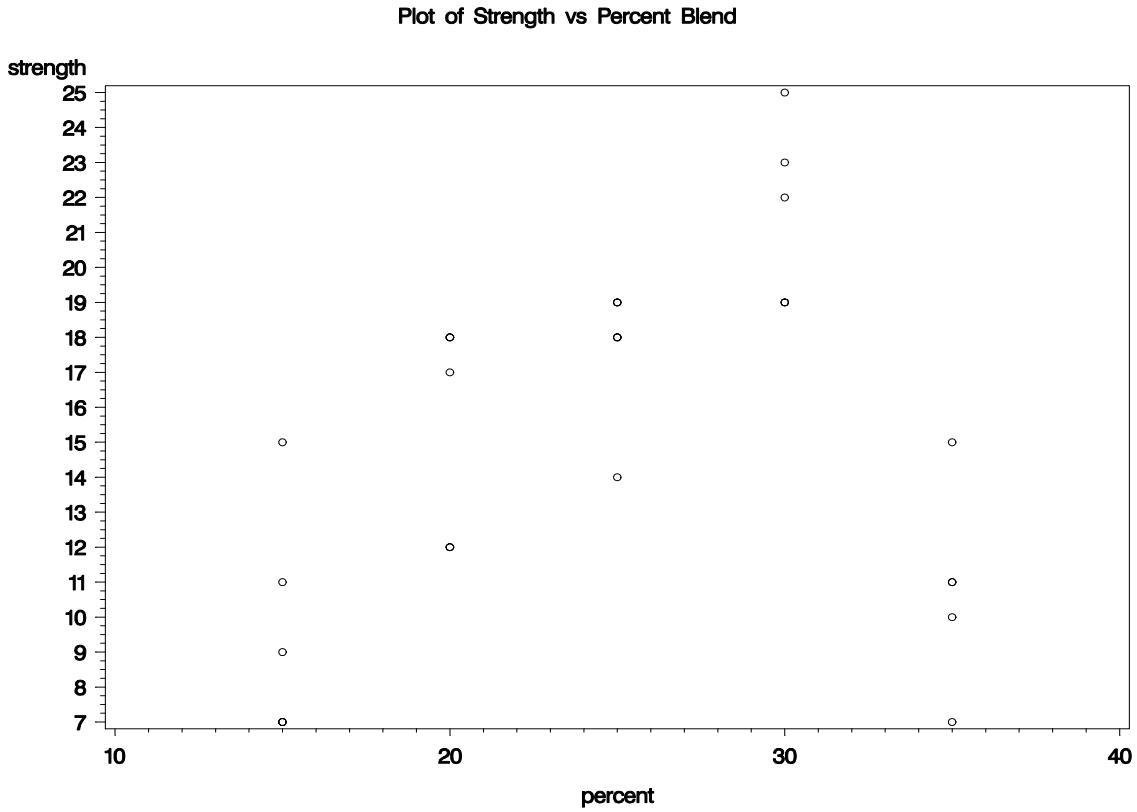
```
options ls=80 ps=60 nocenter;
goptions device=win target=winprtm rotate=landscape ftext=swiss
   hsize=8.0in vsize=6.0in htext=1.5 htitle=1.5 hpos=60 vpos=60
   horigin=0.5in vorigin=0.5in;
data one;
 infile 'c:\saswork\data\tensile.dat';
 input percent strength time;

title1 'Tensile Strength Example';
proc print data=one; run;

symbol1 v=circle i=none;
title1 'Plot of Strength vs Percent Blend';
proc gplot data=one;  plot strength*percent/frame; run;

proc boxplot;
```

```
 plot strength*percent/boxstyle=skeletal pctldef=4;

proc glm;
 class percent; model strength=percent;
 output out=oneres p=pred r=res; run;

proc sort; by pred;
symbol1 v=circle i=sm50; title1 'Residual Plot';
proc gplot;  plot res*pred/frame; run;

proc univariate data=oneres normal;
 var res;  qqplot res / normal (L=1 mu=est sigma=est);
 histogram res / normal; run;

symbol1 v=circle i=none;
title1 'Plot of residuals vs time';
proc gplot; plot res*time / vref=0 vaxis=-6 to 6 by 1;
run;
```

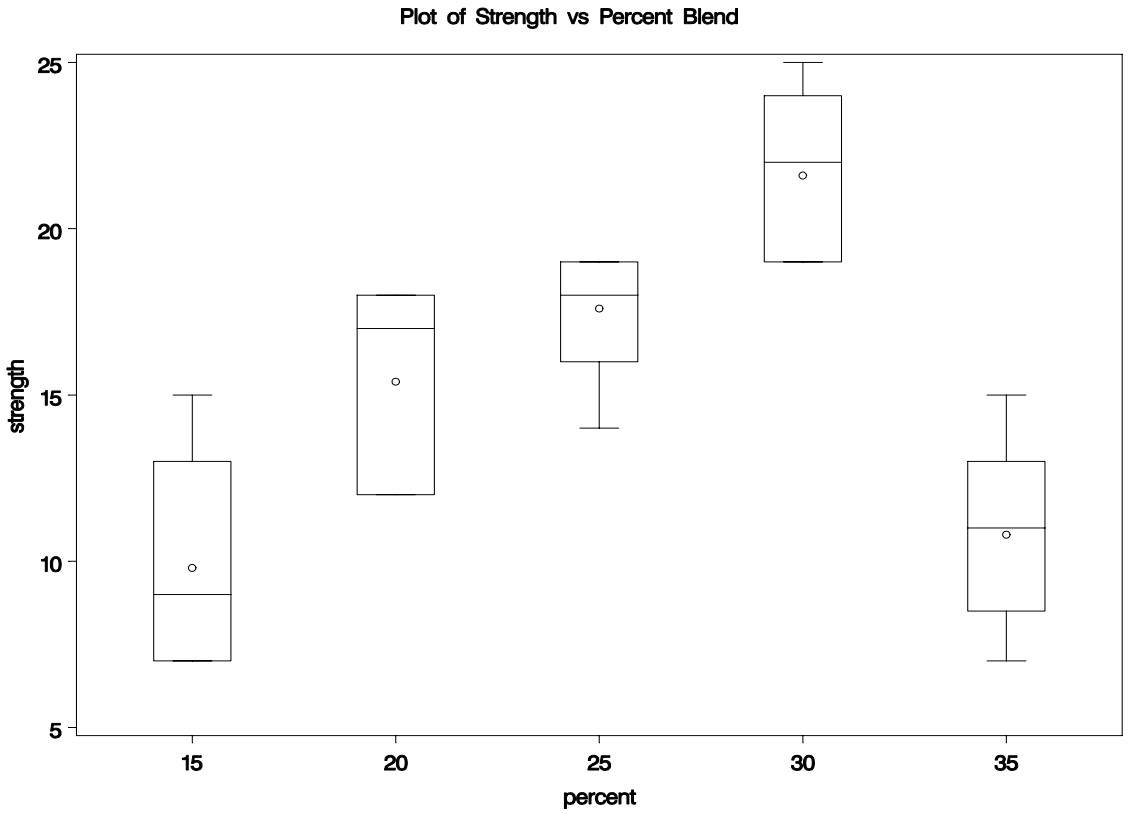
Tensile Strength Example

Obs	percent	strength	time
1	15	7	15
2	15	7	19
3	15	15	25
4	15	11	12
5	15	9	6
6	20	12	8
7	20	17	14
8	20	12	1
9	20	18	11
:	:	:	:
16	30	19	22
17	30	25	5
18	30	22	2
19	30	19	24
20	30	23	10
21	35	7	17
22	35	10	21
23	35	11	4
24	35	15	16
25	35	11	23

Scatter Plot



Side-by-Side Plot



The GLM Procedure
Dependent Variable: strength

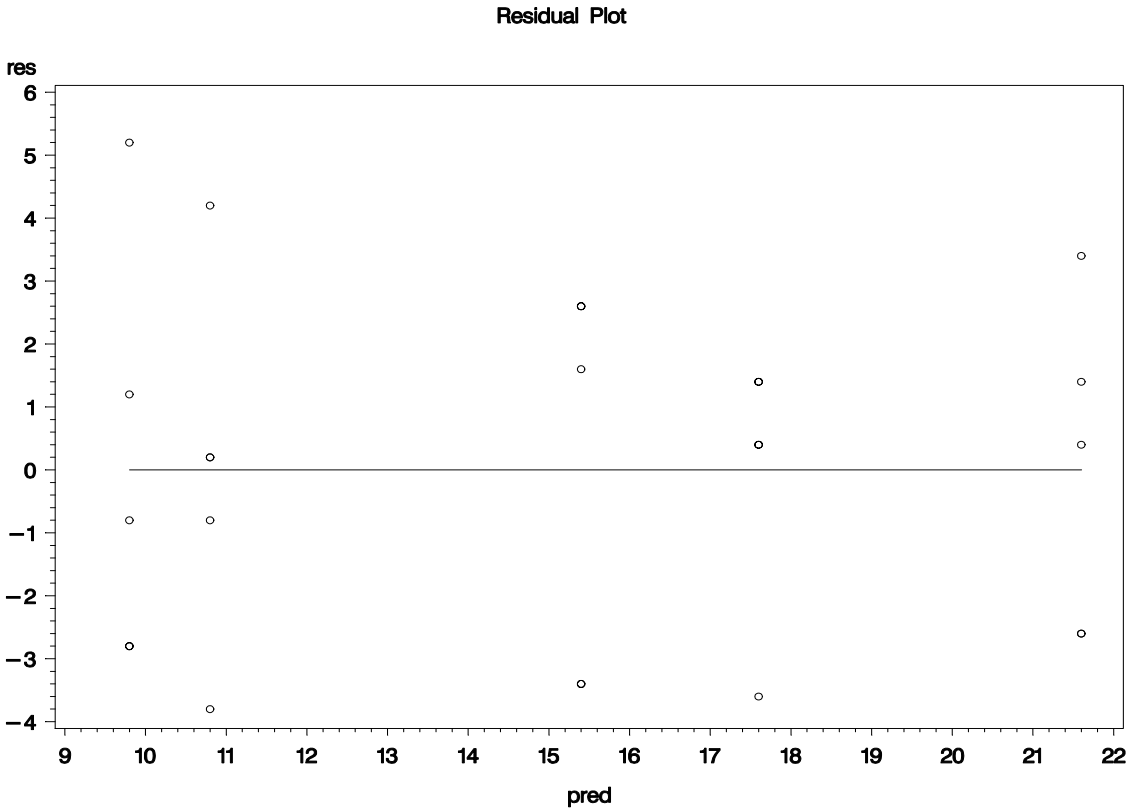
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	475.7600000	118.9400000	14.76	<.0001
Error	20	161.2000000	8.0600000		
Corrected Total	24	636.9600000			

R-Square	Coeff Var	Root MSE	strength Mean
0.746923	18.87642	2.839014	15.04000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
percent	4	475.7600000	118.9400000	14.76	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
percent	4	475.7600000	118.9400000	14.76	<.0001

Residual Plot



The UNIVARIATE Procedure
Variable: res

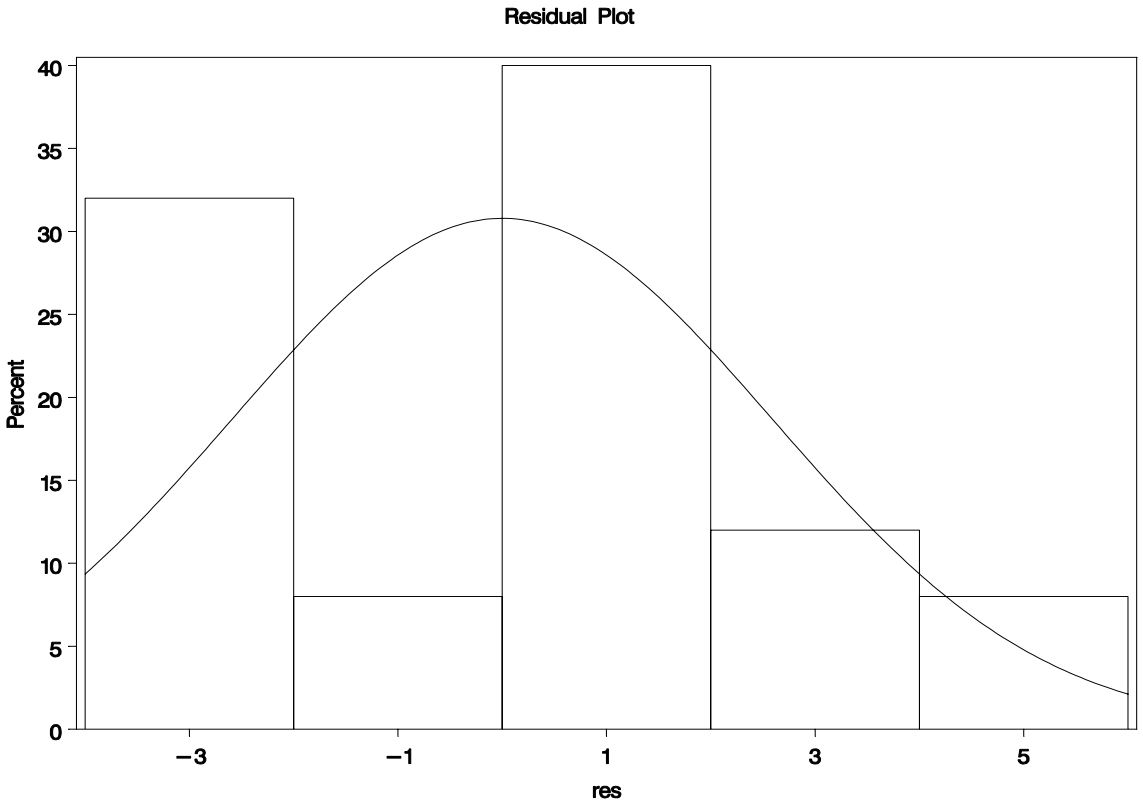
Moments

N	25	Sum Weights	25
Mean	0	Sum Observations	0
Std Deviation	2.59165327	Variance	6.71666667
Skewness	0.11239681	Kurtosis	-0.8683604
Uncorrected SS	161.2	Corrected SS	161.2
Coeff Variation	.	Std Error Mean	0.51833065

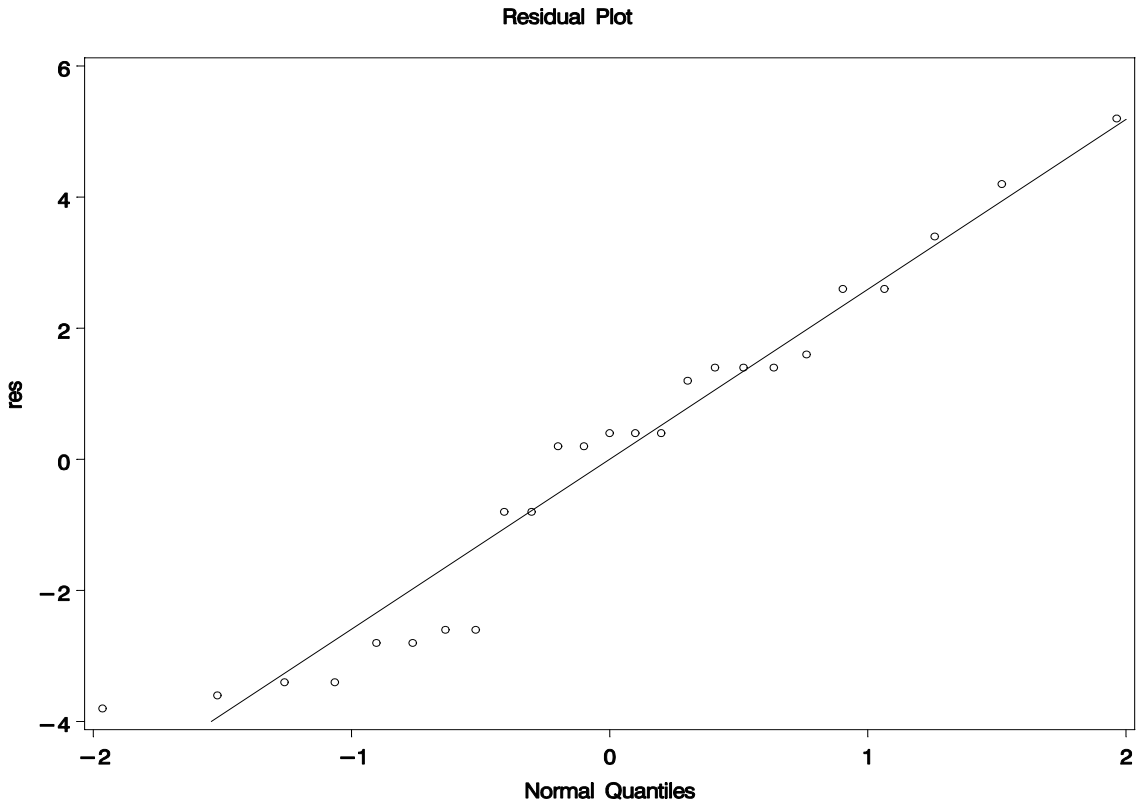
Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.943868	Pr < W	0.1818
Kolmogorov-Smirnov	D	0.162123	Pr > D	0.0885
Cramer-von Mises	W-Sq	0.080455	Pr > W-Sq	0.2026
Anderson-Darling	A-Sq	0.518572	Pr > A-Sq	0.1775

Histogram of Residuals



QQ Plot



Time Serial Plot