1. Machine speed. The number of defective items produced by a machine (Y) is known to be linearly related to the speed setting of the machine (X). The data below were collected from recent quality control records.

a. Fit a linear regression function by ordinary least squares, obtain the residuals, and plot the residuals against X. What does the residual plot suggest?

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: defects

Number of Observations Read	12
Number of Observations Used	12

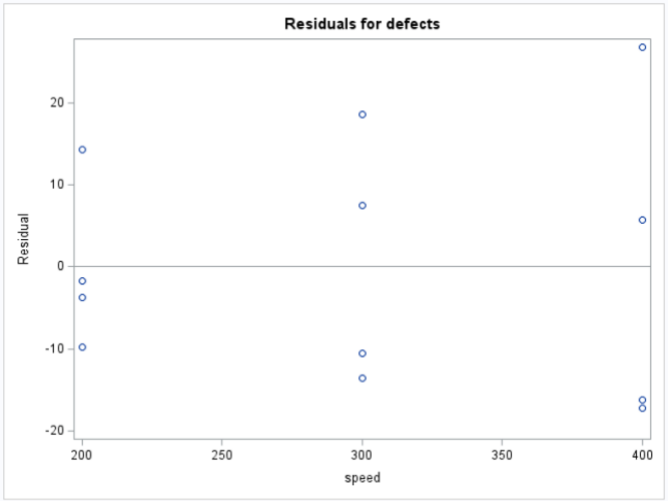
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2812.50000	2812.50000	12.14	0.0059
Error	10	2316.50000	231.65000		
Corrected Total	11	5129.00000			

Root MSE	15.22005	R-Square	0.5484
Dependent Mean	50.50000	Adj R-Sq	0.5032
Coeff Var	30.13872		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-5.75000	16.73052	-0.34	0.7382	-43.02793	31.52793
speed	1	0.18750	0.05381	3.48	0.0059	0.06760	0.30740



The residuals suggest that the assumption of constant variance is violated; residuals clearly increase with greater X values.

b. Conduct the Breusch-Pagan test for constancy of the error variance, assuming $\log_e \sigma^2 = \gamma_0 + \gamma_1 X_i$; use $\alpha = .10$. State the alternatives, decision rule, and conclusion.

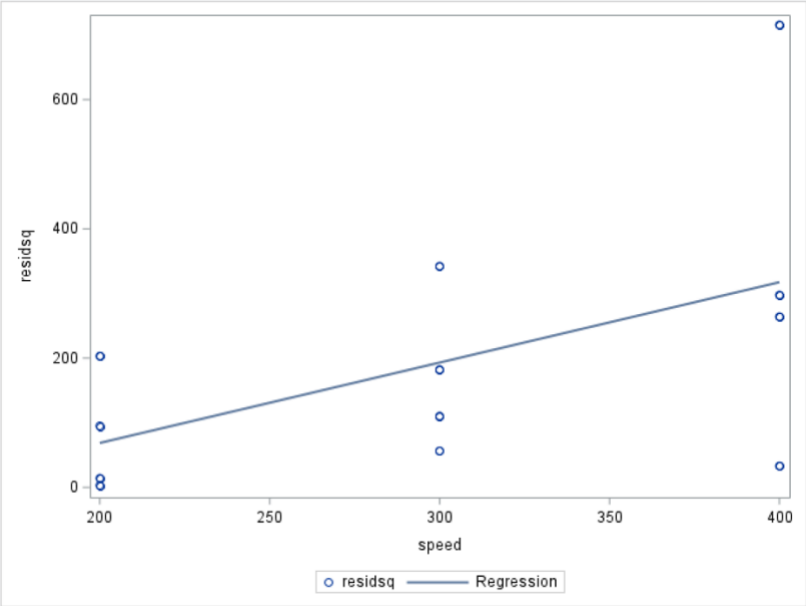
The null and alternative hypothesis are $H_0: \gamma_1 = 0$ and $H_a: \gamma_1 \neq 0$. We reject the null if the p value is less than 0.10.

Nonlinear OLS Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
b0	-5.75	16.7305	-0.34	0.7382
b1	0.1875	0.0538	3.48	0.0059

Since $0.0059 < 0.10$, we reject the null hypothesis and conclude that there is significant evidence to suggest that variance is not constant.

c. Plot the squared residuals against X. What does the plot suggest about the relation between the variance of the error term and X?



There exists a linear relationship between X and the squared residuals, which could suggest that homoscedasticity (error terms have constant variance) is being violated.

d. Estimate the variance function by regressing the squared residuals against X, and then calculate the estimated weight for each case using (11.16b).

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: residsq

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	123753	123753	3.91	0.0762
Error	10	316609	31661		
Corrected Total	11	440362			

Root MSE	177.93503	R-Square	0.2810
Dependent Mean	193.04167	Adj R-Sq	0.2091
Coeff Var	92.17442		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-180.08333	195.59369	-0.92	0.3789
speed	1	1.24375	0.62910	1.98	0.0762

By 11.16b, $w_i = \frac{1}{\hat{v}_i} = (-180.0833 + 1.24375X_i)^{-1}$.

e. Using the estimated weights, obtain the weighted least squares estimates of β_0 and β_1. Are the weighted least squares estimates similar to the ones obtained with ordinary least squares in part (a)?

The REG Procedure
Model: MODEL1
Dependent Variable: defects

Number of Observations Read	12
Number of Observations Used	12

Weight: wt

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	58.95491	58.95491	13.86	0.0040
Error	10	42.53130	4.25313		
Corrected Total	11	101.48621			

Root MSE	2.06231	R-Square	0.5809
Dependent Mean	45.61129	Adj R-Sq	0.5390
Coeff Var	4.52149		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.02847	14.44137	-0.42	0.6852
speed	1	0.18843	0.05061	3.72	0.0040

The weighted least squares estimate using our previously defined weight indeed does yield similar estimates as part (a).

f. Compare the estimated standard deviations of the weighted least squares estimates b_{w0} and b_{w1} in part (e) with those for the ordinary least squares estimates in part (a). What do you find?

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.02847	14.44137	-0.42	0.6852
speed	1	0.18843	0.05061	3.72	0.0040

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.75000	16.73052	-0.34	0.7382
speed	1	0.18750	0.05381	3.48	0.0059

The parameter estimates for WLS and OLS are on the right and left, respectively. We find that the standard error for both the intercept and speed is greater for OLS.

g. Iterate the steps in parts (d) and (e) one more time. Is there a substantial change in the estimated regression coefficients? If so, what should you do?

After reiterating steps (d) and (e), we find no change in the estimated regression coefficients.

2. Refer to Machine speed Problem 11.7. Demonstrate numerically that the weighted least squares estimates obtained in part (e) are identical to those obtained when using transformation (11.23) and ordinary least squares.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
speed_wt	1	0.18843	0.05061	3.72	0.0040
int_wt	1	-6.02847	14.44137	-0.42	0.6852

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.02847	14.44137	-0.42	0.6852
speed	1	0.18843	0.05061	3.72	0.0040

Transformation 11.23 was used to obtain the left while the right was obtained in part (e). They are identical, as desired. The code used to implement the transformation is given below:

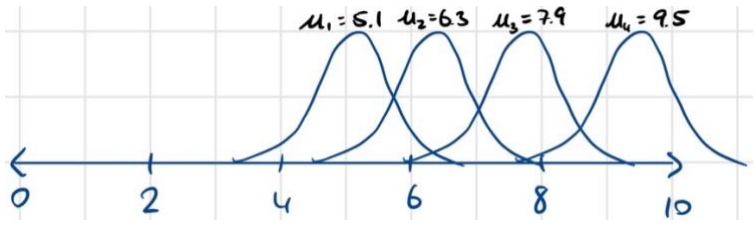
```
DATA a1; INPUT speed defects @@;
CARDS;
200 28 400 75 300 37 400 53 200 22 300 58
300 40 400 96 200 46 400 52 200 30 300 69 .
;

proc reg data=a1 noprint;
   model defects = speed;
   output out=a3 r=resid p=vhat;
run; quit;

Data a3; set a3;
    wt = 1/(vhat);
run; quit;

data a4; set a3;
    defects_wt = sqrt(wt)*defects;
    speed_wt = sqrt(wt)*speed;
    int_wt = sqrt(wt);
run; quit;

proc reg data=a4;
   model defects_wt = speed_wt int_wt / noint;
run; quit;
```

3. A market researcher, having collected data on breakfast cereal expenditures by families with 1, 2, 3, 4, and 5 children living at home, plans to use an ordinary regression model to estimate the mean expenditures at each of these five family size levels. However, the researcher is undecided between fitting a linear or a quadratic regression model, and the data do not give clear evidence in favor of one model or the other. A colleague suggests: "For your purposes you might simply use an ANOVA model." Is this a useful suggestion? Explain.

This is a useful suggestion because an ANOVA model won't assume a prior relationship before testing whether mean expenditure differs by family size, thereby bypassing the issue of choosing between linear versus quadratic regression when there is no conclusive evidence to favor one over the other. Also, since X is a categorical variable (1, 2, 3, 4, or 5 children families), ANOVA will treat each family size as a separate category unlike in regression, which will treat X as continuous.

4. In a study of length of hospital stay (in number of days) of persons in four income groups, the parameters are as follows: $\mu_1 = 5.1$, $\mu_2 = 6.3$, $\mu_3 = 7.9$, $\mu_4 = 9.5$, $\sigma = 2.8$. Assume that ANOVA model, (16.2) is appropriate.

a. Draw a representation of this model in the format of Figure 16.2.



b. Suppose 100 persons from each income group are randomly selected for the study. Find $E\{MSTR\}$ and $E\{MSE\}$. Is $E\{MSTR\}$ substantially larger than $E\{MSE\}$ here? What is the implication of this?

By 16.37a, $E\{MSE\} = \sigma^2 = 2.8^2 = 7.84$. By 16.37b, $E\{MSTR\} = \sigma^2 + \frac{\Sigma n_i(\mu_i - \mu_.)^2}{r-1} = 7.84 + \frac{100(5.1-7.2)^2+100(6.3-7.2)^2+100(7.9-7.2)^2+100(9.5-7.2)^2}{4-1} = 7.84 + \frac{1100}{3} = 374.5067$. $E\{MSTR\}$ is substantially larger than $E\{MSE\}$, so F^* is accordingly very large and so we can conclude that the average length of a hospital stay is not the same across all 4 income categories.

c. If $\mu_2 = 5.6$ and $\mu_3 = 9.0$, everything else remaining the same, what would $E\{MSTR\}$ be? Why is $E\{MSTR\}$ substantially larger here than in part (b) even though the range of the factor level means is the same?

$E\{MSTR\} = 7.84 + \frac{100(5.1-7.3)^2+100(5.6-7.3)^2+100(9-7.3)^2+100(9.5-7.3)^2}{3are} = 7.84 + \frac{1546}{3} = 523.1733$. $E\{MSTR\}$ is substantially larger because compared to their original values, μ_2 and μ_3 are now farther from the grand mean.

5. A completely randomized experiment involving $r = 2$ treatments was carried out, based on $n = 3$ experimental trials for each treatment. The test for equality of the treatment means is to be carried out by means of the randomization distribution of the F^* test statistic (16.55).

a. Determine the number of ways that the six experimental units can be divided into two groups of size three each. How many unique F^* statistics are possible?

There are 6 nCr 3 $= 20$ ways to divide the experimental units under the conditions listed. However, there are only 6 unique F^* values as evidenced in part b:

b. Obtain the randomization distribution of the test statistic F^* and the p-value of the randomization test.

There are 20 ways to permute $\{1, 1, 1, 2, 2, 2\}$ with 1 representing the first treatment and 2 representing the second treatment. These permutations, along with their corresponding F^*-values, are listed below:

1. 111222; 1.64
2. 112122; 1.06
3. 112212; 0.29
4. 112221; 0.59
5. 121122; 0.29
6. 121212; 1.06
7. 121221; 0.59
8. 122112; 1.64
9. 122121; 2.74
10. 122211; 0.97
11. 211122; 0.97
12. 211212; 2.74
13. 211221; 1.64
14. 212112; 0.59
15. 212121; 1.06
16. 212211; 0.29
17. 221112; 0.59
18. 221121; 0.29
19. 221211; 1.06
20. 222111; 1.64

Hence the randomization distribution is given by:

F^* value	2.74	1.64	1.06	0.97	0.59	0.29
$P(F^*)$	0.10	0.20	0.20	0.10	0.20	0.20

The F-value is 1.64 from the observed data. By the randomization test, the associated p-value is $P(F^* = 1.64) + P(F^* = 2.74) = 0.20 + 0.10 = 0.30$.

c. Obtain the p-value of the normal-theory F^* statistic for the sample results in part (b). How does this p-value compare with the one from the randomization test in part (b)? What does this suggest about the appropriateness of the F distribution here if the error terms are far from normally distributed?

The SAS System

The GLM Procedure

Dependent Variable: value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	726.000000	726.000000	1.64	0.2690
Error	4	1766.000000	441.500000		
Corrected Total	5	2492.000000			

R-Square	Coeff Var	Root MSE	value Mean
0.291332	61.79971	21.01190	34.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treatment	1	726.0000000	726.0000000	1.64	0.2690

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treatment	1	726.0000000	726.0000000	1.64	0.2690

The F-value is 1.64 and the associated p-value is 0.2690, which is smaller than the p-value obtained from the randomization test. The F distribution is still appropriate because it does not rely on the assumption that error terms must be normally distributred.

6. Give a table of sample sizes, means, and standard deviations for the six different filling machines.

Level of machine_number	N	deviation	
		Mean	Std Dev
1	20	0.07350000	0.19252546
2	20	0.19050000	0.18540070
3	20	0.46000000	0.16463676
4	20	0.36550000	0.16535369
5	20	0.12500000	0.17273344
6	20	0.15150000	0.17348669

7. Statistically examine the question of whether or not the six machines place the same amount of fill into the cartons. Write a model for this analysis, state the null and alternative hypothesis in terms of your model parameters (cell-based or factor effects), give the test statistic with degrees of freedom, the P-value, and your conclusion.

The SAS System

The GLM Procedure

Dependent Variable: deviation

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.28934667	0.45786933	14.78	<.0001
Error	114	3.53060000	0.03097018		
Corrected Total	119	5.81994667			

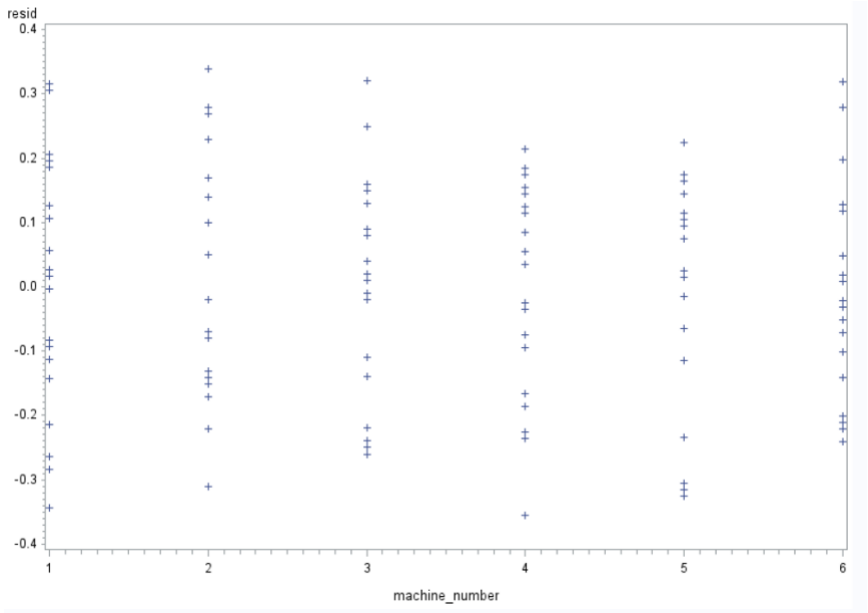
R-Square	Coeff Var	Root MSE	deviation Mean
0.393362	77.29873	0.175983	0.227667

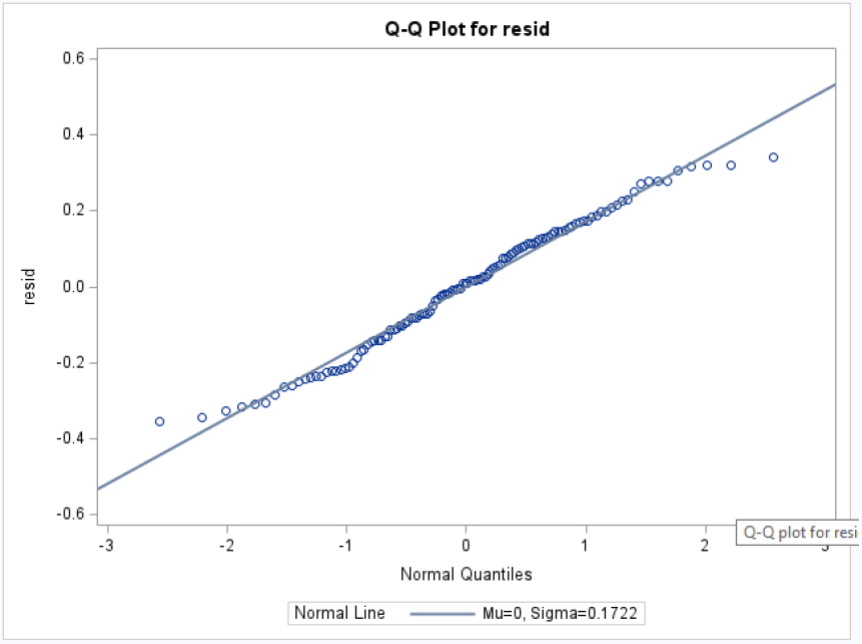
Source	DF	Type I SS	Mean Square	F Value	Pr > F
machine_number	5	2.28934667	0.45786933	14.78	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
machine_number	5	2.28934667	0.45786933	14.78	<.0001

The hypotheses are: H_0: all μ_i, $i = 1,2,\ldots,6$, are equivalent, and H_a: at least one μ_i is different. The test statistic, degrees of freedom, and p-value are all given above. With $p < 0.001$, we reject the null hypothesis and conclude that at least one of the machines has a mean fill amount that significantly deviates from the others.

8. Examine the residuals of this analysis to make sure the assumptions are not violated. Display (and comment on) the plots and/or tests you use to do this.





We examine the residuals by way of a qqplot and a scatterplot of residuals versus machine number. The qqplot indicates that the residuals more or less normally distributed, while the scatterplot indicates there are no outlying observations and that residuals have constant variance.