1. Suppose 90 percent of all batteries from a supplier have acceptable voltages for operation. A random sample of 20 batteries is collected and the batteries are randomly assigned to flashlights. Each flashlight receives two batteries. There are 10 flashlights total.

For a flashlight to be operational, both batteries in it must have acceptable voltages (i.e., each flashlight is a two-component **series system**). Assume that each battery operates independently of other batteries.

(a) Calculate the probability that a single flashlight is operational.

(b) Among the 10 flashlights, what is the probability that at least 9 will be operational?

(c) Redo parts (a) and (b) under the assumption that each flashlight is a two-component **parallel system**; i.e., a flashlight is operational if at least one of its two batteries has an acceptable voltage. Continue to assume that all batteries are independent.

2. Explosive devices used in mining operations produce circular craters when detonated. The radii of these craters Y follow an exponential distribution with $\lambda = 0.2$ meters.

(a) Find the probability that a single crater's radius will be between 5 and 15 meters.

(b) Sketch a graph of the probability density function (pdf) of Y and indicate on the graph the probability you calculated in part (a). Label axes. Neatness counts.

(c) The area of a crater with radius Y is $W = \pi Y^2$. Calculate the expected area $E(W)$.

3. A viticulturist is interested in modeling the time until failure for a new cooling unit specifically designed for wine cellars. He has access to the manufacturer's published data on failure times for $n = 25$ units tested; these data are shown below:

0.32	1.15	1.43	1.47	1.60	1.62	2.12	2.38	2.50	2.82	2.82	2.98	3.00
3.02	3.19	3.44	3.77	3.79	3.89	3.99	4.07	4.10	4.17	4.18	4.19	

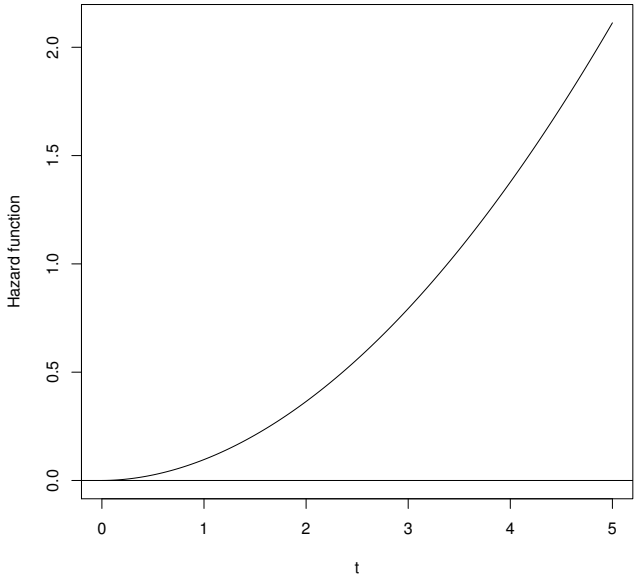
The data are measurements of T, the time (in years) until the cooling unit fails. He assumes a Weibull(β, η) model for T.

(a) Here are the maximum likelihood estimates of the shape (β) and scale (η) parameters based on the data above:

```
> fitdistr(failure.times,densfun="weibull")
  shape    scale
  2.917    3.216
 (0.501)  (0.229)
```

With $\widehat{\beta} \approx 2.917$ and $\widehat{\eta} = 3.216$, estimate $\phi_{0.5}$, the median time until failure.

(b) The quantities in parentheses above (0.501) and (0.229) are the standard errors of the maximum likelihood estimates. Explain to the viticulturist what these measure.



(c) The estimated hazard function from the data is

$$h_T(t) = \frac{2.917}{3.216}\left(\frac{t}{3.216}\right)^{2.917-1},$$

and is shown above. Explain to the viticulturist what information is available from the estimated hazard function $h_T(t)$.

4. The 2008 article, "Development of novel industrial laminated planks in sweetgum lumber" (*Journal of Bridge Engineering*, **13**, 64-66), described the testing of composite beams used to add value to low-grade sweetgum lumber. A random sample of $n = 30$ beams was used. For each beam, investigators recorded Y = breaking strength (measured in lb/in^2). Here are the data:

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7442.69	7886.87	6316.67	7713.65	7503.33	7674.99

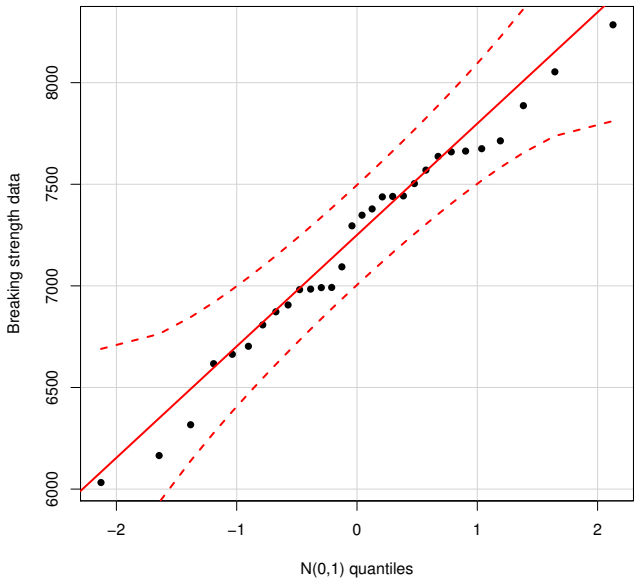
I used R to calculate a 90 percent confidence interval for the population mean beam breaking strength:

```
t.test(strength,conf.level=0.90)$conf.int
[1] 7035.15 7372.56
```

(a) Interpret what this interval means precisely.

(b) If you calculated a 90 percent prediction interval for a 31st beam's breaking strength, would it be shorter than or longer than the interval above? Explain.

(c) Here is the qq plot for the breaking strength data under the normal population distribution assumption:



Suppose an investigator asks you,"Why does a linear trend in the qq plot support the normal distribution assumption? Could the population distribution be something else?" How would you respond to her? To answer her first question, explain precisely how a qq plot is formed and then explain the logic behind how it is interpreted.

5. I recently reviewed a grant proposal for the Hong Kong Research Grants Council. The proposal described an observational study performed last year involving two independent samples of Hong Kong area high school students:

- students who were "non-heavy" smartphone users ($n_1 = 101$)
- students who were "heavy" smartphone users ($n_2 = 103$).

"Heavy" use was defined as using a smartphone ≥ 3 hours/day. A student was called "non-heavy" if s/he was not "heavy."

(a) One variable measured on each student was whether s/he experienced sleep problems or insomnia. There were 8 students in Group 1 ("non-heavy") and 17 students in Group 2 ("heavy") who did. I calculated a 95 confidence interval for the population difference $p_1 - p_2$ ("non-heavy" minus "heavy") using

$$(\widehat{p}_1 - \widehat{p}_2) \pm 1.96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}},$$

and I obtained $(-0.175, 0.003)$.

```
> proportion.diff.interval(8,101,17,103,conf.level=0.95)
[1] -0.175  0.003
```

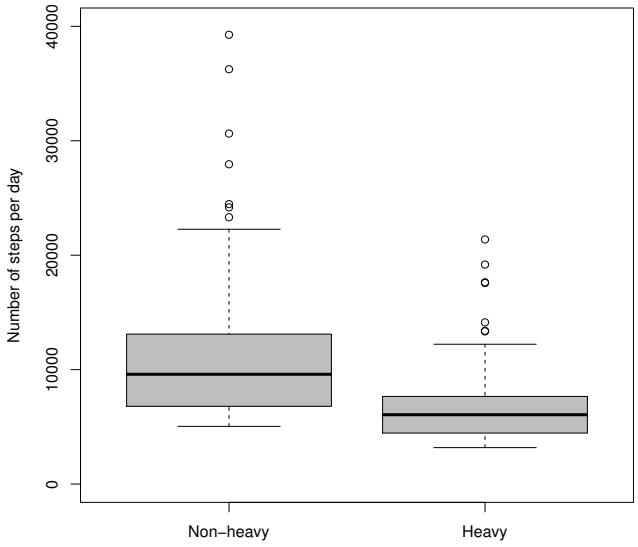
Interpret precisely what this interval means and infer what this suggests about the two groups.

(b) When I changed the confidence level in part (a) to 90 percent, I got this interval:

```
> proportion.diff.interval(8,101,17,103,conf.level=0.90)
[1]  -0.160 -0.011
```

Does it bother you that one interval includes "0," and the other does not? Explain.

(c) Another variable recorded was Y, the number of steps each student took per day (averaged over his/her time in the study). Students wore pedometers each day; these are devices that automatically record this information. Here are side-by-side boxplots of the data collected (again, independent samples; sample sizes given on the previous page):



If the investigators wanted to compare the population mean number of steps per day for these two groups, how would you advise them to do it?

Describe what statistical procedure you would use and how you might check the underlying assumptions associated with this procedure. You don't have to perform any calculations here.

6. An herbal medicine is tested on $n = 16$ randomly selected patients with sleep disorders. Each patient's amount of sleep (in hours) is measured for one night without the herbal medicine and for one night with the herbal medicine. Here are the data from the study:

Patient	Without	With
1	1.8	3.0
2	2.0	3.6
3	3.4	4.0
4	3.5	4.4
5	3.7	4.5
6	3.8	5.2
7	3.9	5.5
8	3.9	5.7
9	4.0	6.2
10	4.9	6.3
11	5.1	6.6
12	5.2	7.8
13	5.0	7.2
14	4.5	6.5
15	4.2	5.6
16	4.7	5.9

(a) Explain why this is a matched pairs study.

(b) I calculated a 95 percent confidence interval for the difference of the population means (without minus with) in R:

```
> t.test(diff,conf.level=0.95)$conf.int
[1] -1.813927 -1.236073
```

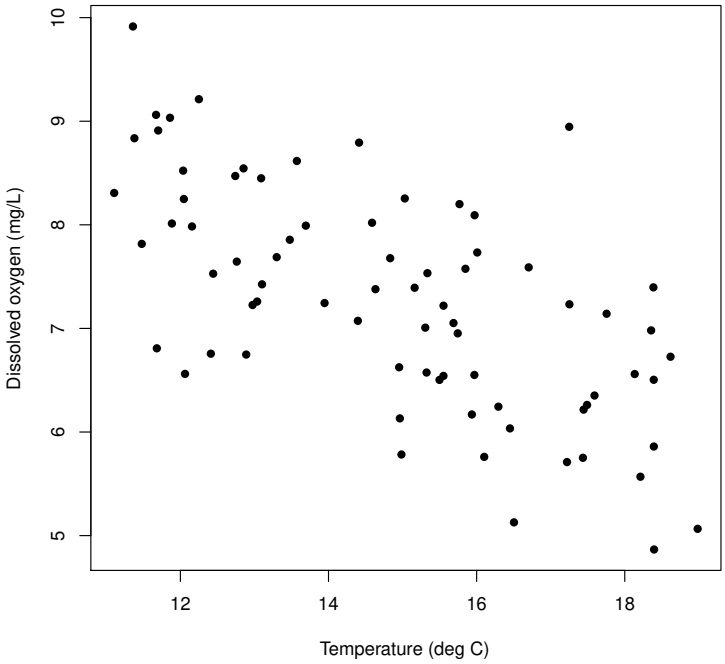
Interpret what this interval means precisely and then carefully describe what effect the herbal medicine has on the amount of sleep.

(c) Suppose you wanted to redesign this study using two independent samples of patients: 16 patients randomly assigned to take the herbal medicine and 16 different patients assigned to take a placebo (a pill designed to look like the medicine, but it contains nothing). Note that this design requires 32 patients total. Do you think the observed data from this study would produce a narrower or wider confidence interval for $\mu_1 - \mu_2$ when compared to the interval in part (b)? Carefully explain your answer.

7. Researchers recorded the following variables on $n = 75$ water specimens taken from a very large lake in northern California:

$$\begin{aligned} Y &= \text{level of dissolved oxygen (mg/L)} \\ x &= \text{water temperature (deg C).} \end{aligned}$$

A scatterplot of the data is shown below:



Consider the population-level model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., 75$, as a model for these data. I fit this model in R and obtained the following output:

```
> fit = lm(dissolved.o2 ~ temp)
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.90532    0.64108  18.571  < 2e-16 ***
temp        -0.31143    0.04261  -7.309 2.78e-10 ***

Residual standard error: 0.827 on 73 degrees of freedom
Multiple R-squared:  0.422
```

I also asked R to calculate 95 percent confidence intervals for β_0 and β_1:

```
> confint(fit)
              2.5 %   97.5 %
(Intercept) 10.6276  13.1829
temp        -0.3963  -0.2265
```

(a) Are temperature and the level of dissolved oxygen linearly related in the population? Explain your answer with statistical evidence.

(b) From the output, we see the value of $R^2 \approx 0.422$ (or 42.2 as a percent). Interpret precisely what this means.

(c) I calculated a 95 percent confidence interval when temperature = 15 deg C:

```
> predict(fit,data.frame(temp=15),level=0.95,interval="confidence")
       fit      lwr     upr
    7.2338   7.0431  7.4245
```

Interpret what this interval means.

8. This problem deals with an extrusion process used in soybeans; basically "extrusion" refers to the process by which certain materials are extracted from the soybeans (e.g., fiber, oil, etc.) to be used in other products (e.g., cattle feed, flour, etc.). An experiment was performed to investigate the relationship between

$$Y = \text{soluble dietary fiber percentage (SDFP) in soybean residue}$$

to three independent variables

- x_1 = extrusion temperature, (`temp`, in deg C)
- x_2 = feed moisture (`moisture`, in %)
- x_3 = extrusion screw speed (`speed`, in rpm).

Here are the data recorded in the experiment:

Observation	x_1	x_2	x_3	Y
1	35	110	160	11.13
2	25	130	180	10.98
3	30	110	180	12.56
4	30	130	200	11.46
5	30	110	180	12.38
6	30	110	180	12.43
7	30	110	180	12.55
8	25	110	160	10.59
9	30	130	160	11.15
10	30	90	200	10.55
11	30	90	160	9.25
12	25	90	180	9.58
13	35	110	200	11.59
14	35	90	180	10.68
15	35	130	180	11.73
16	25	110	200	10.81
17	30	110	180	12.68

Experimenters initially considered the multiple linear regression model to relate `SDFP` to the three independent variables:

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

for $i = 1, 2, ..., 17$.

(a) This regression model can be written out in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. What is the dimension of the design matrix $\mathbf{X}$? Give me the first two rows of the matrix.

(b) Here is the ANOVA table (with sequential sums of squares) for the multiple linear regression model fit:

```
> fit = lm(SDFP~temp+moisture+speed)

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value  Pr(>F)
temp       1  1.2561  1.2561  1.4289 0.25330
moisture   1  3.4585  3.4585  3.9341 0.06885 .
speed      1  0.6555  0.6555  0.7457 0.40350
Residuals 13 11.4281  0.8791

Residual standard error: 0.9376 on 13 degrees of freedom
Multiple R-squared:  0.3197,    Adjusted R-squared:  0.1627
F-statistic: 2.036 on 3 and 13 DF,  p-value: 0.1585
```

What is the conclusion from this analysis? In your answer, cite the value of the overall F-statistic above (and corresponding p-value). Make sure you say precisely what hypotheses the overall F-statistic is testing.

(c) The experimenters also considered a multiple linear regression model with quadratic terms:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \beta_6 x_{i3}^2 + \epsilon_i,$$

for $i = 1, 2, ..., 17$. The three extra independent variables are the squared versions of x_1, x_2, and x_3, respectively. Here is the ANOVA table (with sequential sums of squares) for this multiple linear regression model fit:

```
> fit.quad = lm(SDFP~temp+moisture+speed+temp.sq+moisture.sq+speed.sq)

Analysis of Variance Table
            Df Sum Sq Mean Sq F value    Pr(>F)
temp         1 1.2561  1.2561  24.860 0.0005486 ***
moisture     1 3.4585  3.4585  68.447 8.763e-06 ***
speed        1 0.6555  0.6555  12.973 0.0048330 **
temp.sq      1 2.5869  2.5869  51.197 3.086e-05 ***
moisture.sq  1 5.5393  5.5393 109.629 1.041e-06 ***
speed.sq     1 2.7967  2.7967  55.351 2.211e-05 ***
Residuals   10 0.5053  0.0505

Residual standard error: 0.2248 on 10 degrees of freedom
Multiple R-squared:  0.9699,    Adjusted R-squared:  0.9519
F-statistic: 53.74 on 6 and 10 DF,  p-value: 4.915e-07
```

Which model seems to fit the data better—the model without the quadratic terms or the model with the quadratic terms? Justify your choice with evidence.

9. An engineer is interested in the effects of cutting speed (A), tool geometry (B), and cutting angle (C) on the response variable Y: lifetime (in hours) of a machine tool. Two levels of each factor are chosen, and three replications of a 2^3 factorial experiment are run. Here are the data:

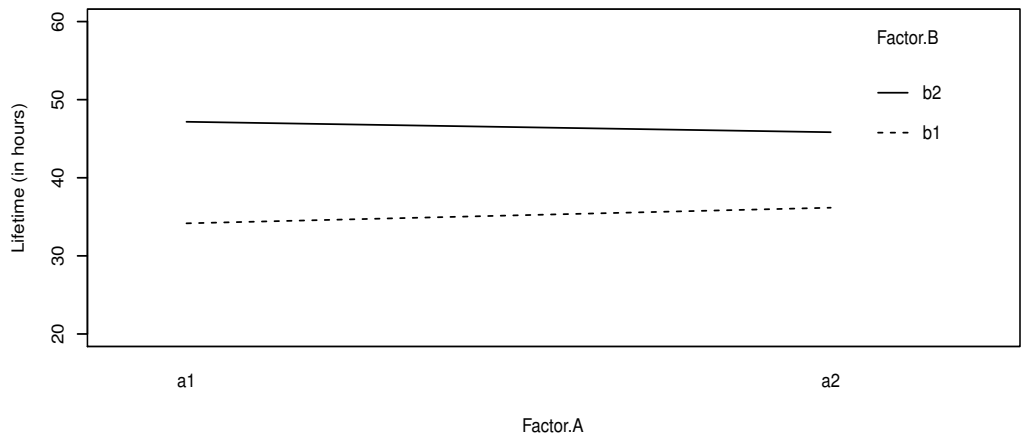
			Treatment	Replicate		
A	B	C	Combination	I	II	III
$-$	$-$	$-$	$a_1b_1c_1$	22	31	25
$+$	$-$	$-$	$a_2b_1c_1$	32	43	29
$-$	$+$	$-$	$a_1b_2c_1$	35	34	50
$+$	$+$	$-$	$a_2b_2c_1$	55	47	46
$-$	$-$	$+$	$a_1b_1c_2$	44	45	38
$+$	$-$	$+$	$a_2b_1c_2$	40	37	36
$-$	$+$	$+$	$a_1b_2c_2$	60	50	54
$+$	$+$	$+$	$a_2b_2c_2$	39	41	47

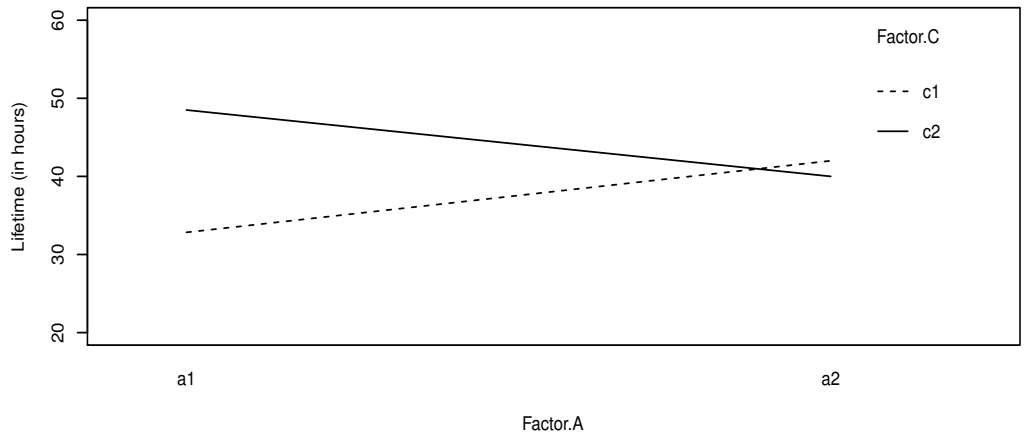
I used R to analyze these data as a 2^3 factorial experiment. The output is below. Recall R's convention; for example, `speed` denotes the main effect of speed; `speed:geometry` denotes the two-way interaction between speed and geometry, etc.

```
> fit = lm(lifetime ~ speed*geometry*angle)
> anova(fit)
Analysis of Variance Table
Response: lifetime
                     Df Sum Sq Mean Sq F value    Pr(>F)
speed                 1   0.67    0.67  0.0221 0.8836803
geometry              1 770.67  770.67 25.5470 0.0001173 ***
angle                 1 280.17  280.17  9.2873 0.0076787 **
speed:geometry        1  16.67   16.67  0.5525 0.4680784
speed:angle           1 468.17  468.17 15.5193 0.0011722 **
geometry:angle        1  48.17   48.17  1.5967 0.2244753
speed:geometry:angle  1  28.17   28.17  0.9337 0.3482825
Residuals            16 482.67   30.17
```

(a) Ignoring two- and three-way interactions, which main effects are significant? Explain.

(b) Examine the two-way interaction plots on the next page. Are all two-way interactions significant? Explain why or why not.





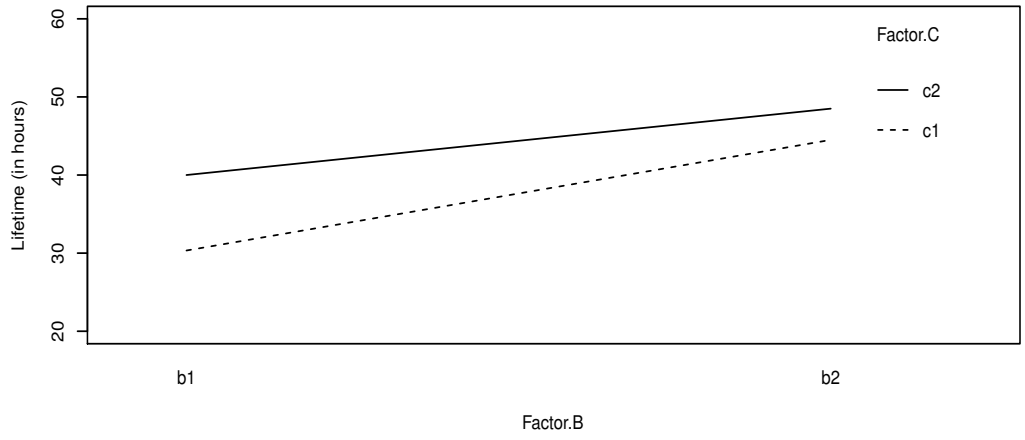


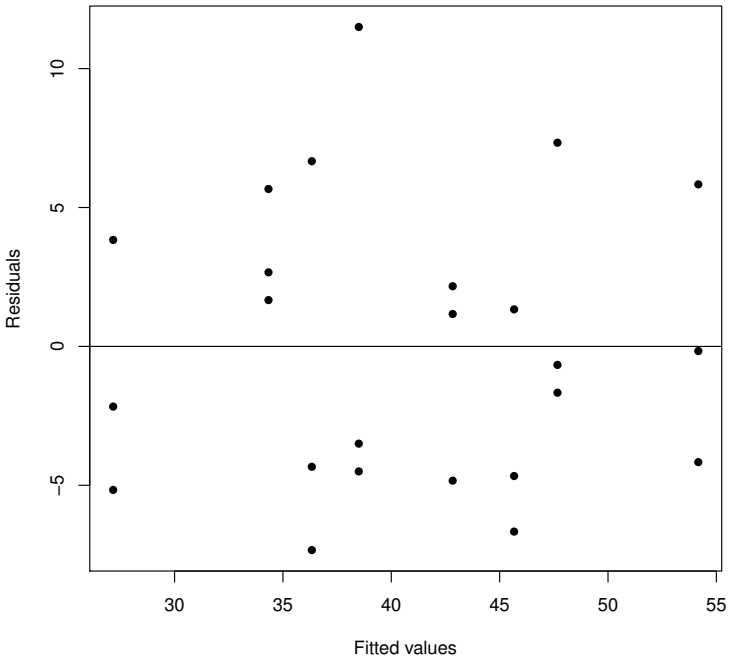
Figure 1: Two-way interaction plots for `speed` (A), `geometry` (B), and `angle` (C).

(c) I refit the model using only the main effects `speed` (A), `geometry` (B), and `angle` (C), and the two-way interaction effect `speed:angle` (AC). Here is the output:

```
> anova(fit)
Analysis of Variance Table
Response: lifetime
            Df Sum Sq Mean Sq F value    Pr(>F)
speed        1   0.67    0.67   0.022 0.8836408
geometry     1 770.67  770.67  25.436 7.216e-05 ***
angle        1 280.17  280.17   9.247 0.0067238 **
speed:angle  1 468.17  468.17  15.452 0.0008972 ***
Residuals   19 575.67   30.30
```

Comparing this ANOVA table to the ANOVA table presented earlier, we see that the residual sum of squares has increased from 482.67 to 575.67. Show precisely where this increase comes from.

(d) After using regression to estimate the model in part (c), I examined the residual plot (i.e., a plot of the residuals versus the fitted values):



What does this plot say about the quality of the model in part (c)?

Binomial:

$$p_Y(y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

Geometric:

$$p_Y(y) = \begin{cases} (1-p)^{y-1} p, & y = 1, 2, 3, ... \\ 0, & \text{otherwise.} \end{cases}$$

Negative binomial:

$$p_Y(y) = \begin{cases} \binom{y-1}{r-1} p^r (1-p)^{y-r}, & y = r, r+1, r+2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Hypergeometric:

$$p_Y(y) = \begin{cases} \dfrac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, & y \leq r \text{ and } n - y \leq N - r \\ 0, & \text{otherwise.} \end{cases}$$

Poisson:

$$p_Y(y) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Exponential:

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases} \qquad F_Y(y) = \begin{cases} 1 - e^{-\lambda y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Gamma:

$$f_Y(y) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Normal (Gaussian):

$$f_Y(y) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/2\sigma^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Weibull:

$$f_T(t) = \begin{cases} \dfrac{\beta}{\eta} \left(\dfrac{t}{\eta}\right)^{\beta-1} e^{-(t/\eta)^\beta}, & t > 0 \\ 0, & \text{otherwise.} \end{cases} \qquad F_T(t) = \begin{cases} 1 - e^{-(t/\eta)^\beta}, & t > 0 \\ 0, & \text{otherwise.} \end{cases}$$