Correlation and Simple Linear Regression

Readings: DVB Ch. 7-9, 27

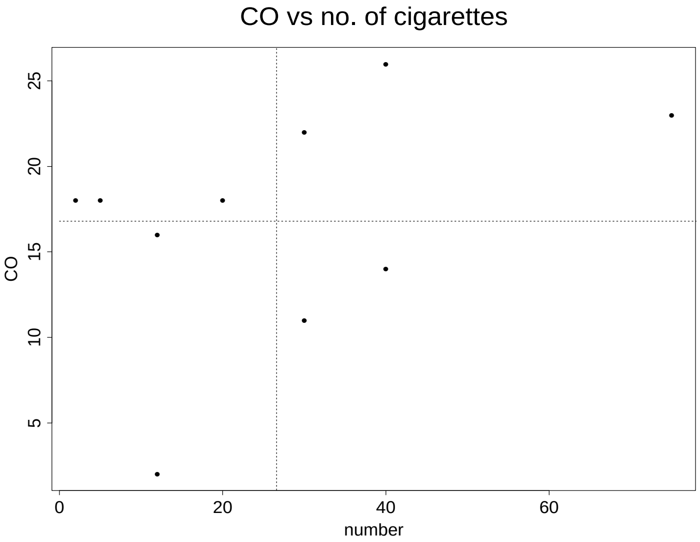
Scatterplot

- Each pair of values is shown as a dot.
- Distance along each axis represents magnitude.
- Reveals nature of association between two continuous random variables.

Example: Carbon monoxide levels were measured in the lung and the 10 subjects were asked how many cigarettes they smoked per day (Selvin, pg. 54)

cigs	12	30	40	12	2	5	20	30	75	40
CO	2	11	14	16	18	18	18	22	23	26

- The plot shows a weak positive association, large CO values tend to occur with large numbers of cigarettes.
- The dashed lines are at the average number of cigarettes and average CO.



- If all the values fell in the bottom left or top right quadrants, the association would be stronger.

Pearson Correlation Coefficient

- Denote the data (x_i, y_i), $i = 1, \ldots, n$, then the Pearson correlation coefficient is

$$r = \frac{\Sigma\,(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma\,(x_i - \bar{x})^2\,\Sigma\,(y_i - \bar{y})^2}}$$

  or

$$r = \frac{1}{n-1}\sum \frac{(x_i - \bar{x})}{s_x}\frac{(y_i - \bar{y})}{s_y}.$$

- r measures the direction and strength of the *linear* association between the two variables.
- It is a dimensionless quantity.
- It has value 1 (-1) if all values on a line with positive (negative) slope.
- It has value 0 if there is no linear association (there may still be a nonlinear association).
- Remember that association does not imply causality!!
- We shouldn't extrapolate relationship beyond range of the data.
- This measure is sensitive to outliers.
- For calculation, use

$$r = \frac{SXY}{\sqrt{SXX \times SYY}}$$

  where

$$SXY = \sum (x_i - \bar{x})(y_i - \bar{y})$$

  or

$$SXY = \sum x_i y_i - (\sum x_i \sum y_i)/n$$
$$SXX = \sum x_i^2 - (\sum x_i)^2/n$$

  and

$$SYY = \sum y_i^2 - (\sum y_i)^2/n.$$

- In the example, with $x =$*number of cigarettes*, $y = CO$, we get

  $$\sum x_i y_i = 5017, \sum x_i = 266, \sum x_i^2 = 11342,$$

  $$\sum y_i = 168, \sum y_i^2 = 3238,$$

  so

  $$SXY = 5017 - 266(168)/10 = 548.2$$

  $$SXX = 11342 - 266^2/10 = 4266.4$$

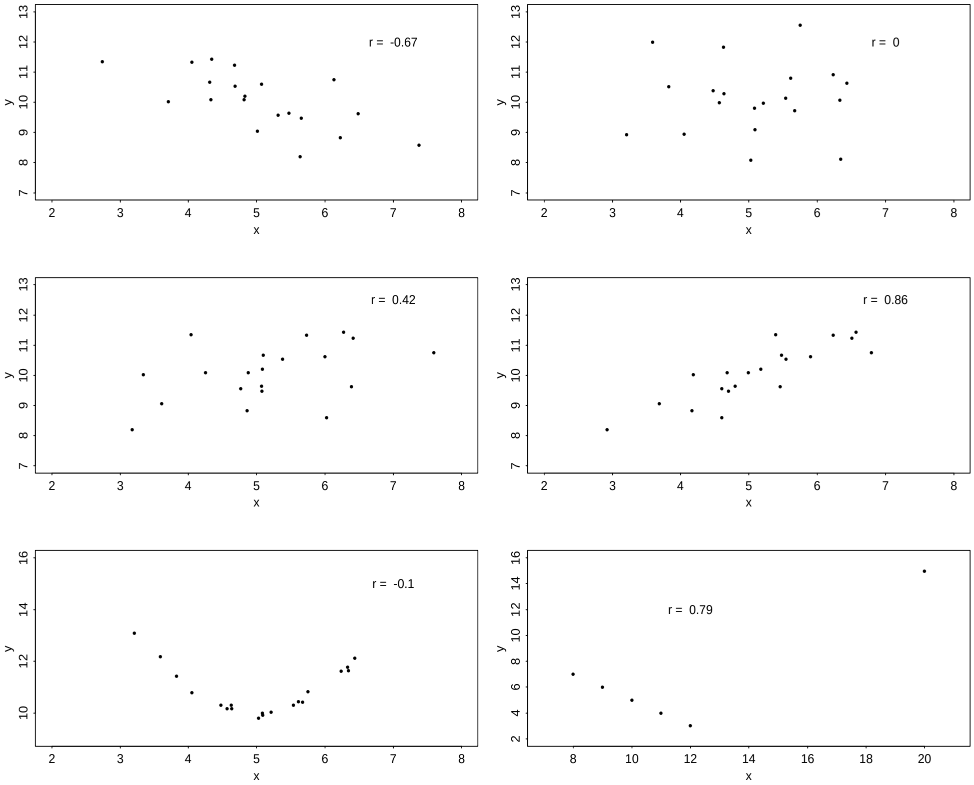
  and

  $$SYY = 3238 - 168^2/10 = 415.6,$$

  so

  $$r = \frac{548.2}{\sqrt{4266.4(415.6)}} = .41$$

- Some examples:



Notes

1. The bottom left example has strong (nonlinear) association, but small r.

2. The bottom right shows reversal of sign due to outlier.

Material on this page concerns test hypotheses regarding the population correlation coefficient ρ. You are not responsible for this material, as one needs some mathematically fairly sophisticated ideas to define the population correlation coefficient. We will see below that testing that the population correlation is 0 is equivalent to testing that the slope of a regression line equals 0, which is a more understandable way of phrasing the problem.

- If our data are a random sample from a population with true correlation ρ, then r is an estimate of ρ with standard error

$$se(\hat{r}) = \sqrt{\frac{1-r^2}{n-2}}.$$

- We can test $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$ using

$$t = \frac{r-0}{se(\hat{r})} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}$$

  which has a t distribution with $n-2$ degrees of freedom if the population is (bivariate) normal.

- In the example,

$$t = \sqrt{10-2}(.41)/\sqrt{1-.41^2} = 1.28$$

  which gives $P = .24$ (using the computer), so there is no evidence that the population correlation is different from 0.

Spearman's (Rank) Correlation Coefficient

- Rank the x's and y's separately.
- Calculate Pearson's correlation coefficient on the ranks.
- In the example, the ranks are as follows (where fractions represent mid ranks)

cigs	3.5	6.5	8.5	3.5	1	2	5	6.5	10	8.5
CO	1	2	3	4	6	6	6	8	9	10

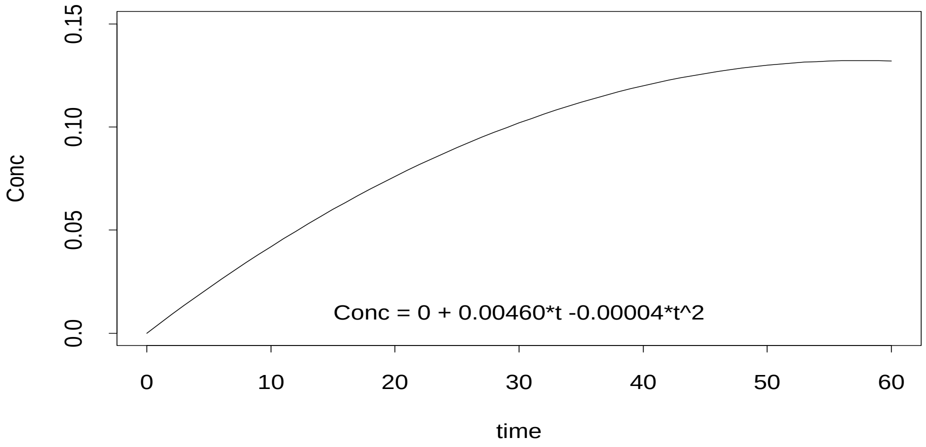
- $r_S = .35$.
- Or we can calculate the difference d_i between the two ranks for each subject

cigs	3.5	6.5	8.5	3.5	1	2	5	6.5	10	8.5
CO	1	2	3	4	6	6	6	8	9	10
d$_i$	2.5	4.5	5.5	-.5	-5	-4	-1	-1.5	1	-1.5

- Then r_S can be calculated as:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

- This gives $r_s = .37$.
- We often get an answer similar to the Pearson correlation coefficient.
- The Spearman correlation is less sensitive to outliers than the Pearson correlation.
- For $n \geq 10$ we can test as before.
- Spearman's measure can be used with ordinal data.
- Spearman's correlation measures *monotone* (possibly nonlinear) association.
- Ranking turns a curved monotone relationship into a linear one.


Conc
0.0
0.05
0.10
0.15
0
10
20
30
40
50
60
time
Conc = 0 + 0.00460*t -0.00004*t^2

Simple Linear Regression

- Simple linear regression is used to find the best straight line fit to the data.
- The result can be used to describe the association, and to predict values.
- We must identify a *response* variable, y, and an *explanatory* variable, x.
- **Statistical Model**

  The observed value y_i includes an additive deviation, ϵ_i

  $$y_i = \alpha + \beta x_i + \epsilon_i.$$

- Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- **Assumptions:** we assume that the deviations $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$. That is, they are independent and have normal distributions with a common mean 0, and common variance σ^2.
- Another way to write the model is

  $$y_i = \mu_{y_i|x_i} + \epsilon_i$$

  where $\mu_{y_i|x_i} = \alpha + \beta x_i$ is the mean of y when the predictor variable x is equal to x_i.

- To find the best line, we minimize the sum of squares of the vertical deviations between the observed and fitted response values at each value of the explanatory variable.
- This is called the *method of least squares*.

- Recall that the equation of a line is

  $$y = a + bx$$

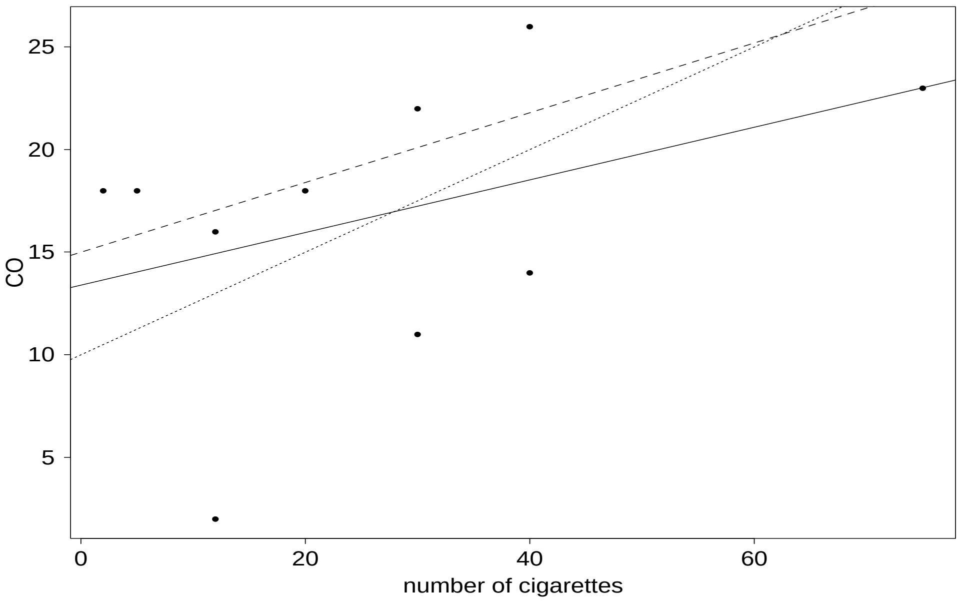
  where a is the $y - intercept$ and b is the *slope*.

- The method of least squares chooses the intercept and slope to minimize the *error sum of squares*

  $$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \alpha - \beta x_i)^2$$

  The minimizing values are denoted as $\hat{\alpha}$ and $\hat{\beta}$.

Example: We would think of CO levels as a response to the number of cigarettes smoked per day.



- The solid line is the least squares fit

$$\hat{y} = 13.38 + .1285x, \quad SSE = 345.16$$

- The other lines are

$$\hat{y} = 10 + .25x, \quad SSE = 408.375$$

and

$$\hat{y} = 15 + .17x, \quad SSE = 426.604.$$

Calculations of the slope and intercept

- The least squares estimates of the slope and intercept can be written as

$$\hat{\beta} = \frac{SXY}{SXX} = r\frac{s_y}{s_x}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

- The slope quantifies the average change in y that corresponds to each one-unit increase in x.
- The intercept, often labeled as the constant in software output, is the expected mean value of y when $x = 0$.
- If x never equals 0, then the intercept has no intrinsic meaning.

Assessing the Fit

- We can summarize the fit using an analysis of variance table.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	70.44	1	70.44
Residual	345.16	8	43.15
Total	415.60	9	

- The Total sum of squares, SST, is the sample variance without the division by $n - 1$

$$SST = \sum(y_i - \bar{y})^2.$$

- The Regression sum of squares, SSR, is obtained by subtraction

$$SSR = SST - SSE.$$

- The total degrees of freedom is $n - 1$.
- The regression degrees of freedom is 1.
- The residual degrees of freedom is the difference, $n - 2$.

The **coefficient of determination**

$$R^2 = \frac{SSR}{SST}$$

is the proportion of variation in y which is explained by x.

- R^2 is between 0 and 1.
- $R^2 = 1$ implies a perfect fit (SSE = 0), all the points are on the line.
- $R^2 = 0$ implies no linear relationship, the best fitting line has zero slope (SSE = SST).
- $R^2 = r^2$, i.e. the coefficient of determination is the square of Pearson's correlation coefficient between x and y.

- In the example $R^2 = 70.44/415.60 = .169$ so 17% of the variation in CO levels can be attributed to variation in numbers of cigarettes smoked.

Hypothesis test for the slope

- Usually we first wish to test whether there is a significant relationship between the variables.
- The hypotheses are $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$
- The test statistic is
$$t = \frac{\hat{\beta} - 0}{\hat{se}(\hat{\beta})}.$$
- This has a t distribution with $n - 2$ degrees of freedom, if assumptions satisfied.
- The standard error of the slope estimate, $\hat{\beta}$, is
$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{SXX}}.$$
- where we estimate σ by
$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}.$$
and $se(\hat{\beta})$ by
$$\hat{se}(\hat{\beta}) = \frac{s}{\sqrt{SXX}}.$$
- Note: this is equivalent to testing $H_0 : \rho = 0$ vs $H_a : \rho \neq 0$.

Confidence Intervals

- We can construct a confidence interval for β, using

$$\hat{\beta} \pm t_{\alpha/2,n-2}\hat{se}(\hat{\beta}).$$

- We typically do the calculations on a computer.
- eg. minitab output for the CO/cigarette data follows

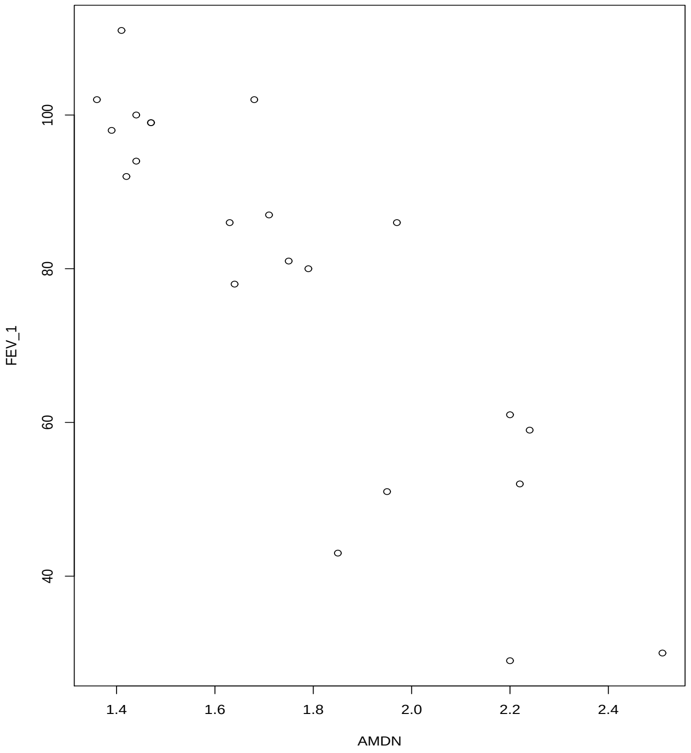
```
MTB > set c1
DATA>  12 30 40 12  2  5 20 30 75 40
DATA> set c2
DATA>  2 11 14 16 18 18 18 22 23 26
DATA> regress c2 1 c1

The regression equation is
C2 = 13.4 + 0.128 C1

Predictor       Coef       Stdev    t-ratio        p
Constant      13.382       3.387       3.95    0.004
C1            0.1285      0.1006       1.28    0.237 *this is the test for
                                                       association, Stdev is
                                                       s/sqrt(SXX)
s = 6.568       R-sq = 16.9%     * s = sqrt(MSE) and estimates sigma
                                 * R^2 is the coefficient of determination
                                     and equals SSR/SST

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1       70.44       70.44      1.63    0.237
Error         8      345.16       43.15
Total         9      415.60

Unusual Observations                       *you can ignore this part
Obs.      C1         C2       Fit Stdev.Fit  Residual   St.Resid
  1     12.0       2.00     14.92      2.54    -12.92     -2.13R
  9     75.0      23.00     23.02      5.29     -0.02     -0.00 X

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

- In this example the test for the relationship between amount of smoking and CO is not significant ($t = 1.28, P = .237$).

Example: (Daniel) Habib and Lutchen present a diagnositic technique that is of interest to repiratory disorder specialists. The following are the scores elicited by this technique, called AMDN, and the forced expiratory volume (FEV_1) scores for 22 subjects.

AMDN	FEV_1
1.36	102
1.42	92
1.41	111
1.44	94
1.47	99
1.39	98
1.47	99
1.79	80
1.71	87
1.44	100
1.63	86
1.68	102
1.75	81
1.95	51
1.64	78
2.22	52
1.85	43
2.24	59
2.51	30
2.20	61
2.20	29
1.97	86

A plot of the data shows a strong negative association.



Some data summaries are $n = 22$, $\Sigma x_i = 38.74$, $\Sigma y_i = 1720$, $\Sigma x_i y_i = 2875.01$, $\Sigma x_i^2 = 70.6468$, $\Sigma y_i^2 = 147138$.

- From these we can calculate

$$\bar{x} = \sum x_i / n = 1.76, \quad \bar{y} = 1720/22 = 78.18,$$

$$\begin{aligned} SXX &= \sum x_i^2 - (\sum x_i)^2/n \\ &= 70.6468 - (38.74)^2/22 = 2.4292, \end{aligned}$$

$$\begin{aligned} SXY &= \sum x_i y_i - \sum x_i \sum y_i/n \\ &= 2875.01 - 38.74(1720)/22 \\ &= -153.7536 \end{aligned}$$

and

$$\begin{aligned} SYY &= \sum y_i^2 - (\sum y_i)^2/n \\ &= 147138 - 1720^2/22 \\ &= 12665.27 \end{aligned}$$

- Pearson's correlation coefficient is

$$\begin{aligned} r &= \frac{SXY}{\sqrt{SXX \times SYY}} = \frac{-153.7536}{\sqrt{2.4292(12665.27)}} \\ &= -.8766 \end{aligned}$$

- The least squares estimate of slope is

$$\hat{\beta} = \frac{SXY}{SXX} = \frac{-153.7536}{2.4292} = -63.3.$$

- The intercept estimate is

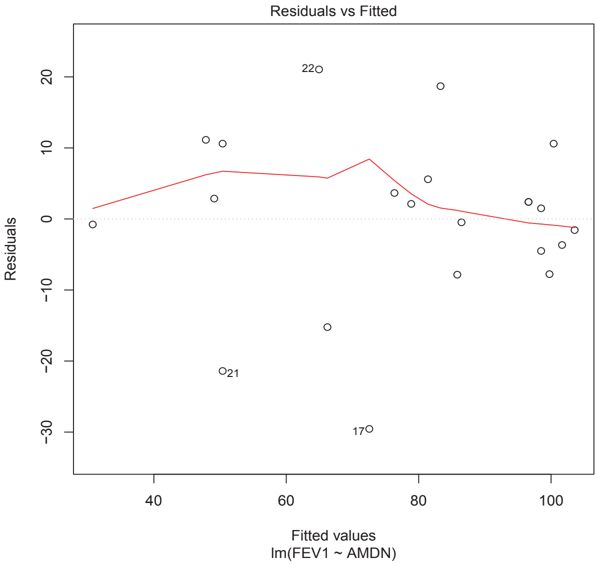
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 78.18 + 63.3(1.76) = 189.6$$

- Some computer output follows from the R package

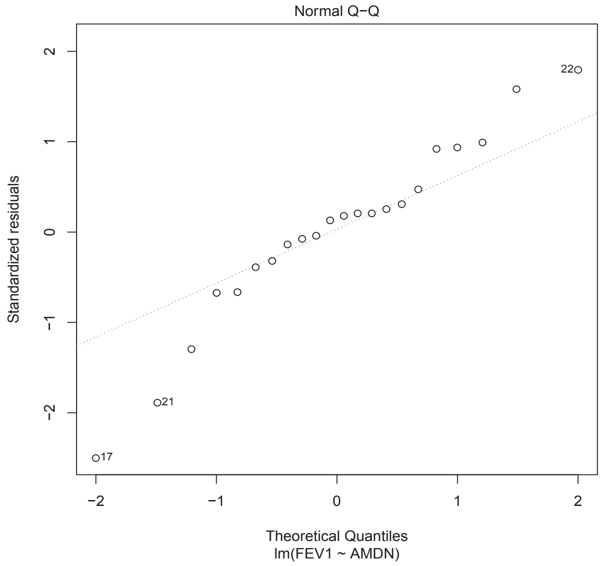
```
> summary(resp.out)

Call:
lm(formula = resp.y ~ resp.x)

Residuals:
    Min      1Q  Median      3Q     Max
-29.543  -4.285   1.817   5.112  21.052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  189.638     13.925  13.619 1.41e-11 ***
resp.x       -63.294      7.771  -8.145 8.81e-08 *** #this is the test for
                                                        association
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 12.11 on 20 degrees of freedom *this is s=sqrt(MSE)
Multiple R-squared: 0.7684,
F-statistic: 66.35 on 1 and 20 DF,  p-value: 8.814e-08
> anova(resp.out)
Analysis of Variance Table        *unlike MINITAB, this program doesn't give
                                   the SST

Response: resp.y
          Df Sum Sq Mean Sq F value    Pr(>F)
resp.x     1 9731.7  9731.7  66.348 8.814e-08 ***
Residuals 20 2933.5   146.7
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Residual Plots - used to check assumptions for linear regression

- Assumptions: Residuals are independent and have normal distributions with a common mean 0, and common variance σ^2.
- A plot of the residuals $\epsilon_i = y_i - \hat{y}_i$ versus x or ϵ_i versus $\hat{y}_i$, the fitted or predicted values of the response variables, should show random scatter.
- Curvature indicates a more complicated model is required, such as a quadratic $y = a + bx + cx^2 + \epsilon$.
- A 'fanning out' of residuals indicates the variance changes with the mean or with x.
- Outliers in the x or y variable could show up.



- These residuals show no problems in the randomness of the residuals. Subjects 17, 21 and 22 have the largest absolute values of the residuals.
- Another often used residual plot is the quantile-quantile (Q-Q) plot. The following figure shows the normality of the residuals is somewhat violated especially for those subjects having large absolute residual values.



Note: you are not responsible for this page.

The confidence interval for $\mu_{y|x}$, the mean of y when the predictor variable equals x, is

$$\hat{\alpha} + \hat{\beta}x \pm t_{\alpha/2,n-2}\hat{se}(\mu_{y|x}).$$

where

$$\hat{se}(\mu_{y|x}) = s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

A **prediction interval** for a new y at x is

$$\hat{\alpha} + \hat{\beta}x \pm t_{\alpha/2}^{n-2}\hat{se}(\hat{y}_{new})$$

where

$$\hat{se}(\hat{y}_{new}) = s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

Because $\hat{se}(\hat{y}_{new})$ is larger than $\hat{se}(\mu_{y|x})$, the prediction interval for a new y at x is wider than the confidence interval for $\mu_{y|x}$.