RNR / ENTO 613 --Assumptions for Simple Linear Regression

Statistical statements (hypothesis tests and CI estimation) with least squares estimates depends on 4 assumptions:

1. Linearity of the mean responses
2. Constant variance of the responses around the straight line
3. Normality of subpopulations (Y's) at the different X values
4. Independence within and across subpopulations

Relative importance of these assumptions depends on type of inferences:

Proposed Inference	Required Assumptions
Predicting a single Y at X; or single X at Y (prediction interval)	All assumptions of models are necessary (normality is assumed)
Estimating the mean Y at X; or mean X at Y (confidence interval).	All assumptions except normality (normality of distribution of mean is met because of Central Limit Theorem)
Estimating value of the slope	Only linearity maters (lack of independence affects the SE but not the estimate of the slope).

Model Assessment with graphical tools: Scatterplot of response variable versus Explanatory variable

A simple scatterplot of Y * X is useful to evaluate compliance to the assumptions of the linear regression model. The pattern for the means and variability of the responses suggests a strategy for analysis.

<<Fig. 8.6 Sleuth>>

1- The regression is a straight line and the variability of Y is about the same at all locations along the line: use simple linear regression.
2- Regression is not a straight line but SD is constant: transform X
3- Regression line is not monotonic: use multiple regression.
4- Regression not a straight line and SD increase as a function of X: transform Y
5- Regression is a straight line, homogenous SD, but outliers: use simple regression tool and report presence of outliers
6- Regression is a straight line but SD increases in X: use weighted regression.

Note on Weighted regression:

The variance may increase with increases in the explanatory variable even if the regression line is linear. This suggests that the chance fluctuation (i.e. variance of estimation error) affecting Y increases with increased values of X.

For example, on days when many people drive to the park, many other people may also get there by other means. Thus the more people in the park, the greater would be the error about the regression line: SD (Y) would be proportional to X.

With such a pattern, it is not recommended to predict single or average values of Y or X with a simple linear regression approach because no single estimate of SE applies across all the Xs values. However, the slope of the line can be estimated without bias using *weighted regression.*

Weighted regression gives less influence to values of Y that are subject to larger error for estimating the parameters of the line. In weighted regression, the least squares estimate minimizes:

$$\sum_{i=1}^{n} w_i \left(Y_i - \hat{Y}_i\right)^2$$

where w_i is a weight given to every observation.

Depending on the pattern of the relation between X and the variance of Y, one may choose $w_i = 1 / X_i$ (variance proportional to X: "V" shape of residuals), or $w_i = 1 / X_i^2$ (variance proportional to SQRT X: "Horn" shape of residuals).

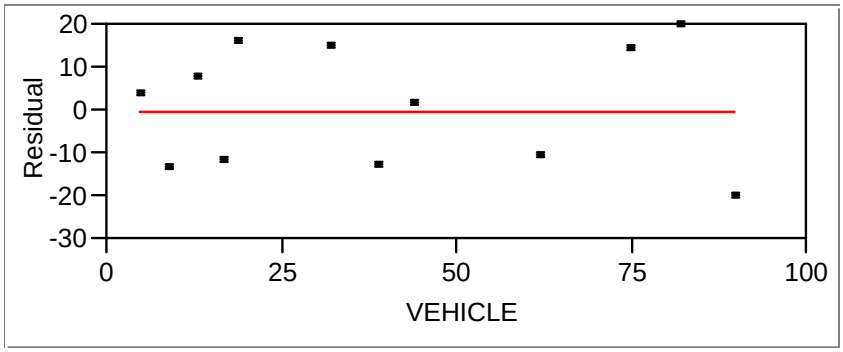
JMP allows computation of weighted regression models (in the Fit Model platform).

Scatterplot of residuals versus Fitted values

Assessment of the adequacy of regression models is done with plots of *residual* vs *fitted values*. These show the error (residual) about the fitted line when the linear component of variation has been removed. Looking at distribution of the residuals about a horizontal line makes it easier to assess curvature and spread of observations.

These plots are essential for evaluating *nonlinearity*, *nonconstant variance*, and presence of *outliers*.

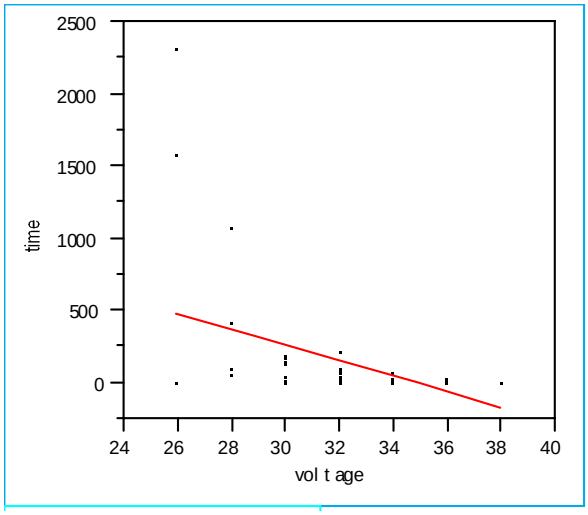
The Park example:



Assessing whether a transformation improves fit of the model is done by trial and error.

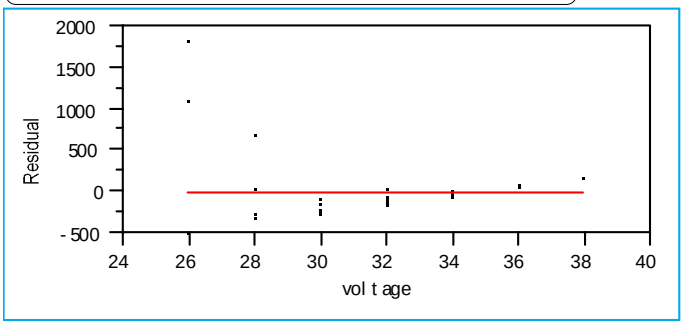
Example: breakdown times for insulating fluids under different voltages (Sleuth case 8.1.2)

Time = 1886.17 – 53.95 Voltage

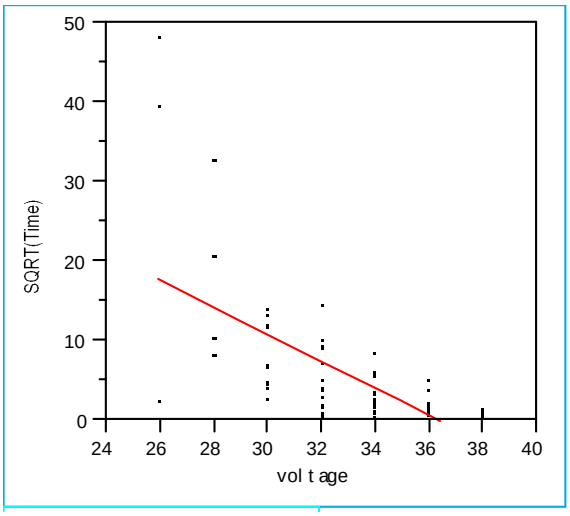


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1886.1695	364.4812	5.17	<.0001
voltage	-53.95492	10.95264	-4.93	<.0001

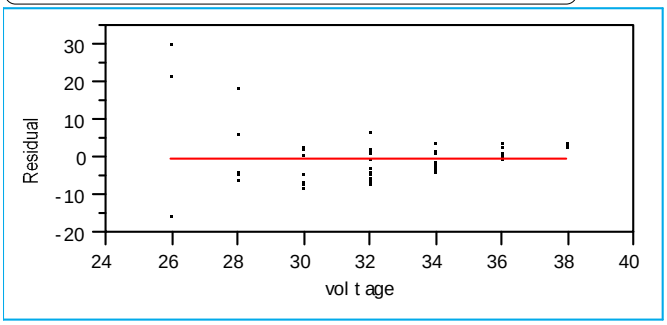


SQRT(Time) = 61.78 – 1.69 Voltage

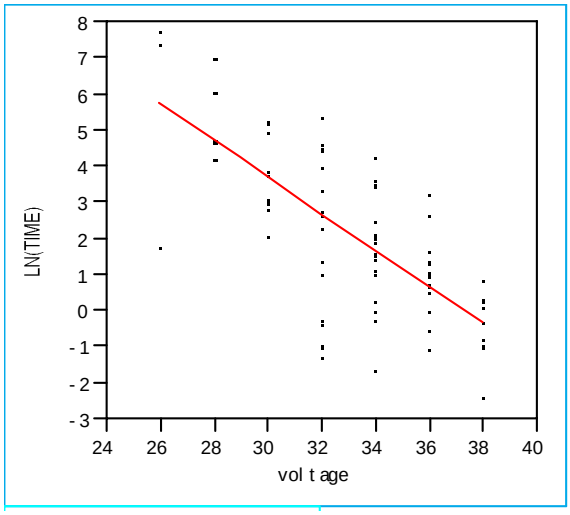


Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	61.784472	7.776881	7.94	<.0001
voltage	-1.695897	0.233695	-7.26	<.0001

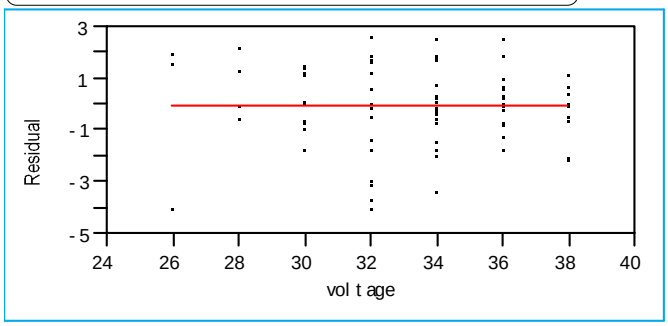


Ln (Time) = 18.95 – 0.50 Voltage



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18.955459	1.910019	9.92	<.0001
voltage	-0.507365	0.057396	-8.84	<.0001



Testing whether the regression parameters differ from zero is not assessing model fit. In the above examples, the null hypothesis of $\beta_1 = 0$ is rejected *in the three analyses*, but the assumption of linearity and homogeneity of variance are only met when a logarithmic transformation of Y is used.

Interpretation of linear relationship after log transformation

Only Y is log transformed:

A *one-unit change in X* results in a multiplicative change of *exp* (β1*)* in the *median of Y*:

$$\text{Median}\{Y \mid (X+1)\} / \text{Median}\{Y \mid X\} = \exp(\beta 1)$$

Example: Breakdown time versus Voltage of insulation fluids

Ln (Time) = 18.95 – 0.50 (Voltage). Each increase of 1 kV results in a change of $e^{(-0.50)}$ = 0.61 in the median breakdown time. Thus each increase in 1 kV results in a 0.61 decrease (i.e., 39%) in breakdown time.

The 95 % CI for β1 was –0.62 to –0.39, so the 95 % CI for the f0.61 decrease is exp (-0.62) to exp (-0.39), or 0.54 to 0.68.

Only X is log transformed:

A *doubling of* X results in a *β1 log (2)* change in the *mean* of Y:

$$\mu\{Y \mid \ln(2X)\} - \mu\{Y \mid \ln(X)\} = \beta 1 \ln(2)$$

Example: Change in pH as a function of time in meat

pH = 6.98 – 0.726 ln (Time). A doubling of time after slaughter is associated with a ln (2)(-0.726) = 0.503 unit change in pH, i.e. the mean pH is reduced by 0.503 for each doubling of time after slaughter.

The 95% CI for β1was from –0.805 to –0.646, so the 95 % CI for the reduction in pH per doubling of X is from ln (2)(-.805) to ln (2) (-.646), or from 0.449 to 0.558.

Both X and Y are log transformed:

A *doubling of* X results in a multiplicative change of $2^{\beta 1}$ in the *median* of Y

$$\text{Median}\{Y \mid \ln(2X)\} - \text{Median}\{Y \mid \ln(X)\} = 2^{(\beta 1)}$$

Example: Island size versus number of species

Log (Species) = 1.94 + 0.250 log (Area). Thus an island of area 2A will have a median number of species that is $2^{0.250}$ greater (i.e. 1.19 fold) than an island of area A. The 95% CI for β1 was from 0.219 to 0.281, so the 95 % CI for the change in median is $2^{0.219}$ to $2^{0.281}$, or 1.16 to 1.22. << Display 8.2>>

Interpretation for other types of transformations

May be difficult. Interpretation is not required:

1) if the regression is used to detect presence of an *association* between two variables.
2) to predict values of X and Y. Here the only requirement is to back transform the predicted value (and their CI).

Simple regression analysis with a F-test (extra-sum-of-squares F-test)

The approach for linear regression described so far used the distribution of *the least squares estimates* and t-tools to draw statistical inferences.

Another approach uses the *Extra sum of squares method* to compare the regression model (full model) to the grand mean model (reduced model). To test the linear model, we compare the difference in "explanatory power" between the reduced and a full model.

The sum of the *residual sum of squares* measures the *variability* in the observations that *remains unexplained* after we fit a regression model.

The sum of squares (SS) calculated after fitting the Grand mean is:

$$\text{Total SS} = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2$$

The SS calculated after fitting the regression line is:

$$\text{Error SS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

The difference between the two is the *extra* variability explained by the regression model:

Total SS	–	Error SS	=	Model SS
Unexplained by Grand mean	–	Unexplained by regression		Extra variation explained by regression
df = n – 1		df = n –2		df = 1 = [n –1] – [n –2]

The model comparison (regression vs grand mean) is given by:

$$F = \frac{Model\ SS\ /\ df}{Error\ SS/\ df}$$

If the null hypothesis of *equal* mean responses is correct (i.e. $\beta_1 = 0$), then the numerator and denominator of the F-statistic both estimate the same population variance (σ^2), so the F-statistic should be close to one. If the null hypothesis is not correct, the F-statistic will be greater than 1.

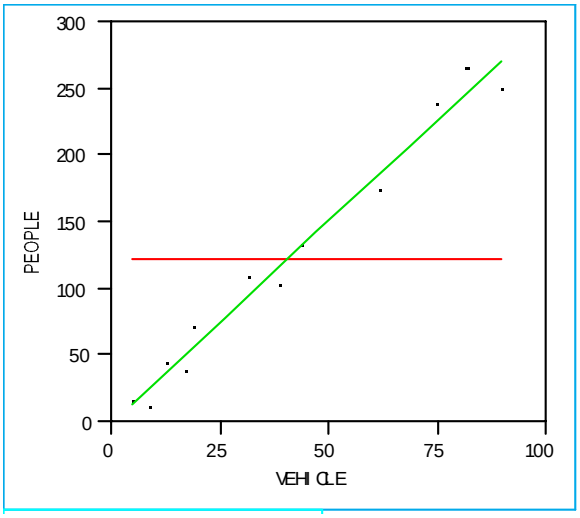
The Park example: n = 12 observation pairs.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	90688.256	90688.3	418.9975
Error	10	2164.411	216.4	Prob>F
C Total	11	92852.667		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.947634	7.342907	-0.13	0.8999
VEHICLE	3.0212969	0.1476	20.47	<.0001



(The F-statistic is equal to the square of the t-statistic, i.e. $20.47^2 = 419.00$)

Assessment of the Regression Fit using ANOVA (Lack of fit test):

When a response variable *is measured repeatedly for each (or some) level of the explanatory variable* (i.e. with replicates), the Extra SS approach provides a way for comparing the fit of *simple linear regression model* (reduced model) to the *separate-means model* (full model, i.e. one-way ANOVA).

<<Display 8.4>>

The question: Is the straight-line regression model *sufficient* to describe the data, or should we instead use a separate-mean model? (because the mean responses are not linear)

For i groups, we can fit the following *hierarchical* models:

1. Separate-means model (Anova):	$\mu\{Y\mid Xi\}= \mu i$	(i parameters)
2. Simple linear regression model:	$\mu\{Y\mid Xi\}= \beta o + \beta 1 Xi$	(2 parameters)
3. Equal-means model:	$\mu\{Y\mid Xi\}= \mu$	(1 parameter)

We can now use the *Extra sum of squares method* to compare the simple regression and ANOVA models (i.e., perform a lack of fit test).

Example: Insulation fluid data (7 Voltage treatments)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	196.47741	32.7462	13.0043
Error	69	173.74892	2.5181	Prob>F
C Total	75	370.22633	4.9364	<.0001

ANOVA (X coded as ordinal)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	190.15149	190.151	78.1409
Error	74	180.07484	2.433	Prob>F
C Total	75	370.22633		<.0001

REGRESSION (X coded as continuous)

In both cases, Total SS is the variability left unexplained by fitting a single mean, while Error SS is the variability left unexplained by each model. Error SS is smaller for the ANOVA than for the regression because the separate means explain sightly more variation than the regression line (but it uses more parameters).

Lack-of–Fit F-test

We compare the fit of the separates means (ANOVA) to the fit of the regresion line, by comparing the residual SS from both models using the *Extra sum of squares F-test*:

$$F = \frac{(SS\ Res_{LR} - SS\ Res_{SM})/(df_{LR} - df_{SM})}{\sigma^2_{SM}}$$

Where LR and SM stands for linear regression and separate mean model.

Insulating fluid example:

SS Res_{LR} =180.07, SS Res_{SM} = 173.75, df_{LR} =74, df_{SM} = 69, σ^2_{SM} =2.518

$F_{5,69} = [(180.074 - 173.748) / (74 - 69)] / 2.518 = 0.502$, p = 0.78

This large p-value, corresponding to a F ≈ 1, provides no evidence of a lack-of – fit of the simple linear model. A small p-value would suggest that the ANOVA model fits better, or in other words that the variability between group means is not explained adequatedly by a simple linear regression model.

In JMP, you get the lack of fit test both on the *Fit Model platform* and the *Fit Y by X* platform, as long as there are replicated Y's at some levels of X's.

For the Insulation fluid data (7 Voltage treatments):

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	5	6.32592	1.26518	0.5024
Pure Error	69	**173.74892**	2.51810	Prob > F
Total Error	74	**180.07484**		0.7734
				Max RSq
				0.5307

Pure error is the SS for the ANOVA model (i.e. the best estimate of the population variance), whereas total error is the SS for the simple regression model. The lack of fit test (above) assesses the null hypothesis: both models fit equally well, versus the alternative hypothesis: ANOVA fits better.

A composite Analysis of variance table

A composite ANOVA table can be used to sumarize results from the lack of fit test.

The separate-means model (i parameters, here i = 7) is a generalization of the simple linear model (2 parameters), which in turn is a generalization of the equal-mean model (1 parameter). The reduction in SS obtained when fitting the separate-mean model instead of the equal-mean model (i.e. 196.477), can be decomposed in:

1- The reduction in SS obtained when fitting a simple linear regression instead of an equal-mean model (190.151), and
2- The reduction in SS obtained when a separate-means model is fitted instead of a linear regression model (196.477 – 190.151 = 6.326 = Lack of fit SS).

Each of those components can be represented in a composite ANOVA table, which is the same table used to represent the separate means-model, except that the *Between group SS* (treatment SS) is decomposed into two components:

1. One that represents the variability explained by the linear regression line (190.15) , and
2. One that estimates the variability that arizes because the separate group means do not exactly fall on a straight line (the lack-of-fit component, i.e. 6.32).

<<Fig. 8.9 Sleuth>>

R^2: The proportion of variation explained

R^2 is the *coefficient of determination*, a measure of the percentage of the total response variation that is explained by the explanatory variable. Thus,

$$R^2 = 100\left(\frac{Total\ SS - Residual\ SS}{Total\ SS}\right)\% = 100\left(\frac{Model\ SS}{Total\ SS}\right)\%$$

The proportion (or %) of the total variation in Y that is explained by the fitted regression (or by any model) is a measure of the strength of the fitted relationship.

In the insulating fluids example analysed with a linear regression, R^2 = 100(370.2 – 180.1 / 370.2)% = 51.4 %.

Thus, *Fifty-one percent of the variation in log breackdown times was explained by the linear regression on voltage.*

R^2 provides useful information to compare the fit of 2 models with **equal** numbers of parameters. R^2 alone should not be used to assess adequacy of the linear model because R^2 can be large even when the linear model is not adequate (e.g., the response is not linear).

Simple Linear regression or One-way ANOVA?

When data are collected from groups that correspond to different levels of an explanatory variable (eg. insulation fluids), both one-way ANOVA and simple linear regression can be used for analysis.

The choice between the 2 is easy:

If the simple linear regression fits, then it is preferred.

When appropriate, regression is better than ANOVA because it:

1. allows for interpolation
2. provides more degrees of freedom for error estimation (thus small error MS and high power)
3. gives smaller SE for estimates of mean responses (i.e., narower CI).

<<Display 8.11 Sleuth>>

4. provides a simpler model

In general, we seek the simplest model --the one with the fewest parameters—that adequately fits the data. This is called the principle of *parsimony*, or the *Occam's* (or Ockham) *razor* principle.