Model Checking

Recall: For both the ANOVA model as well as the linear regression model an important model assumption is that the residuals ϵ_i are independent, normally distributed with mean zero and common variance σ^2.

If this assumption is not satisfied, it is still possible to conduct the statistical analysis (and SPSS will provide you with a fitted model, and p-values, and confidence intervals...) However, the results will be meaningless!!! That means for instance that the computed p-value is *not* the probability to see more extreme data if the null hypothesis were true. The confidence interval will *not* contain the true population parameter for $(1-\alpha)\%$ of samples etc. Therefore it is extremely important to check these model assumptions before drawing any kind of conclusions from the data.

Residuals: From a Statistician's perspective, the observed residuals ϵ_i are the parts of the data which cannot be explained by the model (everything in life that we don't understand goes into the residuals...) Y_i is the variable you observe, $\hat{Y}_i$ is what you would expect Y_i to be if the model were a perfect fit, thus $\epsilon_i = Y_i - \hat{Y}_i$.

Even though we assume the residuals to have constant variance, since the observed residuals ϵ_i are a function of the data (x_i, y_i), they are themselves random variables and their variance can be computed as

$$V(\epsilon_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

By the definition of the linear regression model, the mean of the residuals $E(\epsilon_i)$ is always equal to zero.

Standardized Residuals: Standardize the residuals, by subtracting the mean (zero) and dividing by the standard deviation estimate (different for each residual):

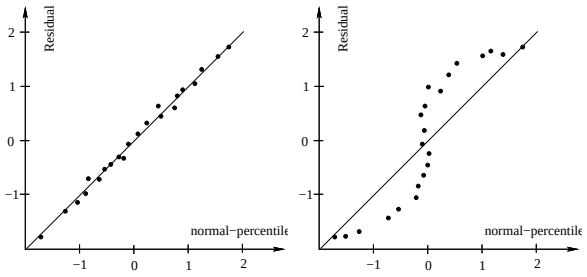
$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}, \qquad i = 1, 2, \ldots, n$$

SPSS calls these the studentized residuals. You can obtain them for ANOVA or regression analysis under the SAVE option tab.

Diagnostic Plots: After fitting a candidate model to the data, the residuals for this model can be computed. Diagnostic plots help to decide whether the residuals conform to the assumptions, which means that the model is a good fit. They can also be used as a guide for modifications if the assumptions are not (yet) satisfied. Iterate this procedure until a model is found that fits the data well.

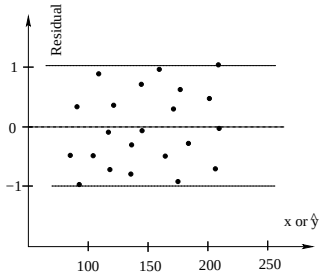
The following list provides suggestions for diagnostic plots and what you should look out for:

Probability Plots are used to check the normality assumption for the residuals. If the dots fall close to the 45° degree line (left graph), we conclude that the residuals have (approximately) a Normal distribution.

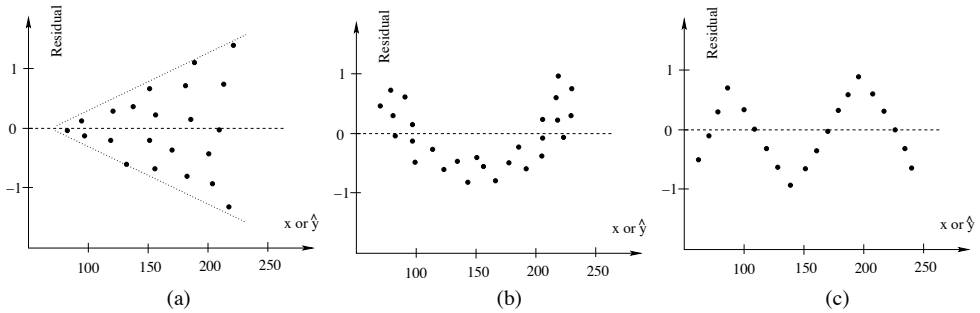


If the dots in a qq-plot or in a pp-plot exhibit any kind of shape other than a straight line (right graph), it suggests that the distribution of the ϵ_i is not normal. In this case, consider a variable transformation (on either your response or the predictor).

A **Residual Plot** is a scatterplot of the residuals. Most commonly, the residuals (on the y-axis) are plotted against either the observed predictor values x_i or the expected response values $\hat{y}_i$. Ideally, we would like to see the dots in a residual plot in a rectangular region with no discernible pattern.



Reasons for concern: A triangular pattern (a) indicates that the residuals do not have constant variance (consider a model transformation or WEIGHTED LEAST SQUARES techniques). Outliers in a residual plot (not shown below) can be used to "flag" influential observations in the data set. Any pattern (b) in a residual plot is an indication that the residuals are a function of the predictor or response (consider fitting y as a nonlinear function of x). A strong pattern (c) in a residual plot indicates that the residuals are not independent of each other.



Variable Transformations

If a residual analysis suggests that there may be a nonlinear relationship between the predictor variable X and the response Y, the model can be modified to achieve a better fit.

DEFINITION: A function relating y to x is INTRINSICALLY LINEAR if by means of a transformation on x and/or y, the function can be expressed as

$$y' = \beta_0 + \beta_1 x'$$

where y' and x' are the respective transformed variables.

DEFINITION: A probabilistic model relating Y to x is INTRINSICALLY LINEAR, if by means of a transformation on Y and/or x, it can be reduced to a linear probabilistic model

$$Y' = \beta_0 + \beta_1 x' + \epsilon$$

The most commonly used variable transformations are:

Function	Transformation(s) to Linearize	Linear Form
Exponential: $y = \alpha e^{\beta x}$	$y' = \ln y$	$y' = \ln(\alpha) + \beta x$
Power: $y = \alpha x^{\beta}$	$y' = \log(y), x' = \log(x)$	$y' = \log(\alpha) + \beta x'$
$y = \alpha + \beta \log(x)$	$x' = \log(x)$	$y = \alpha + \beta x'$
Reciprocal: $y = \alpha + \beta \frac{1}{x}$	$x' = \frac{1}{x}$	$y = \alpha + \beta x'$

Which function transformation will provide the best fit for your particular data set, is more often than not a TRIAL AND ERROR procedure: Fit a model. Check the residuals. If the fit is good, you're done. If the fit is not good, transform one (or both) of the variables. Fit the new model. Check the residuals. If the fit is good, you're done. If not, try again...

Example: PCB concentrations in lake trout.

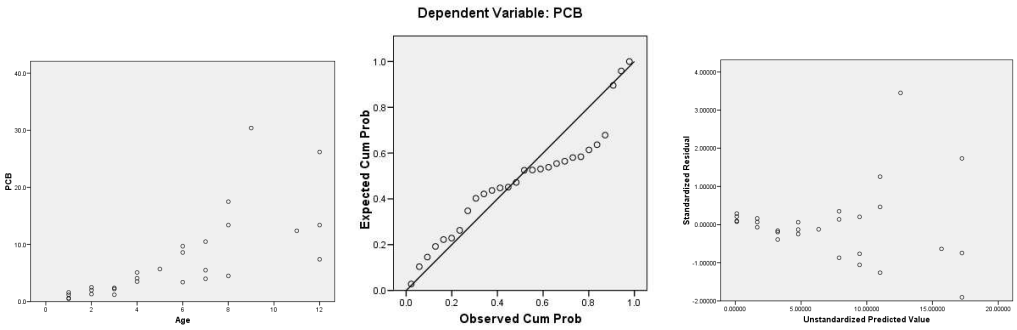
Researchers measured concentration of polychlorinated biphenyl (PCB) residues in a series of lake trout from Cayuga Lake, NY. The ages of the fish were accurately known, because the fish were annually stocked as yearlings and distinctly marked as to the year they were put into the lake. Each whole fish was mechanically chopped, ground, and thoroughly mixed, and 5-gram samples taken (compare "Trout.txt").

Our first attempt is to fit a simple linear regression model to the data

$$\text{PCB} = \beta_0 + \beta_1 \text{ Age} + \epsilon$$

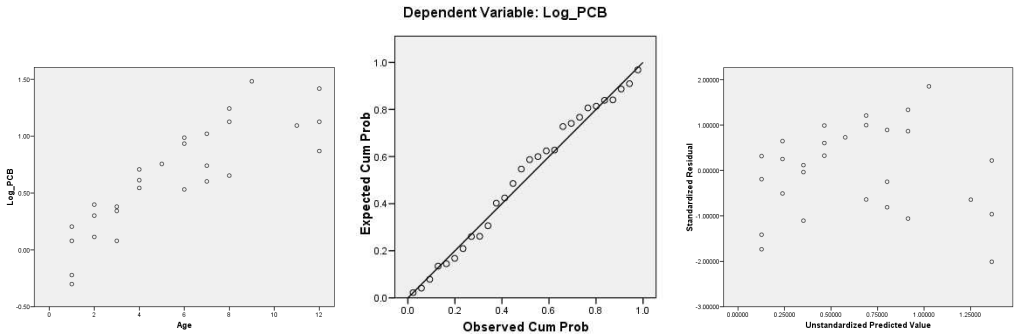
The results of this model can be found below (from left to right: scatterplot of PCB against Age, pp-plot for the residuals, residual plot).

Untransformed model: $Y = \beta_0 + \beta_1 X$ (SPSS: $\hat{\beta}_0 = -1.452, \hat{\beta}_1 = 1.558, R^2 = 0.542$)



(a) Comment on the appearance of the plots. Do you think the simple linear regression model (with untransformed variables PCB and AGE) provides a good fit?

Transformed model: $\log(Y) = \beta_0 + \beta_1 X$ (SPSS: $\hat{\beta}_0 = 0.014, \hat{\beta}_1 = 0.113, R^2 = 0.731$)



(b) Comment on the appearance of these plots. What has changed for this different model? Would you conclude that this model provides a good fit for the data?

Note: The researchers who published this data, concluded that the model $\log(Y) = \beta_0 + \beta_1 X^{\frac{1}{3}} + \epsilon$ provides an even better fit.

Note: Beware of fitting too complicated models. Make sure that the final model you choose makes (some) sense in the field that the data is taken from. While a model such as

$$\ln(Y) = \beta_0 + \beta_1 X + \epsilon$$

is easy to defend (exponential growth rates are very common in many fields) a model such as

$$\tan^{-1}\left(Y + \frac{3}{7}\right) = \beta_0 + \beta_1 \frac{1}{\sqrt{x^2 - 1}} + \epsilon$$

is much harder to explain.

Try to achieve a balance between a high R^2-value, good residual plots and simplicity of the final model. Sometimes a simpler model, even if it has lower R^2 and not as pretty residual plots is preferable, if it can be easily explained in the context the data is taken from.

Logistic Regression

In some linear regression models, the (not so quantitative) response Y is a variable which takes on only two possible values (success/failure, yes/no, 0/1). Let the success probability be

$$p = P(Y = 1).$$

This value of p may depend on the value of the quantitative predictor X, $p = p(x)$.

We could make $p(x)$ a linear function of x

$$p(x) = \beta_0 + \beta_1 x$$

but then we may run into trouble, because p is a probability ($\in [0, 1]$) but the right side of the equation has no such restriction.

Better: Relate p and x via the LOGIT FUNCTION

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Logistic regression means to assume that $p = p(x)$ have the above relationship and to find appropriate values for β_0 and β_1. Straightforward algebra shows that

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \Leftrightarrow \frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x} \Leftrightarrow \ln\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Example: The *Challenger* disaster in January 1986 was caused by the failure of an O-ring. The incidence could have been prevented, because data on failure of these O-rings (as a function of outside air temperature) was available at the time of the shuttle launch.

Temperature	Failure	Temperature	Failure	Temperature	Failure
53	Yes	68	No	75	No
56	Yes	69	No	75	Yes
57	Yes	70	No	76	No
⋮	⋮	⋮	⋮	⋮	⋮

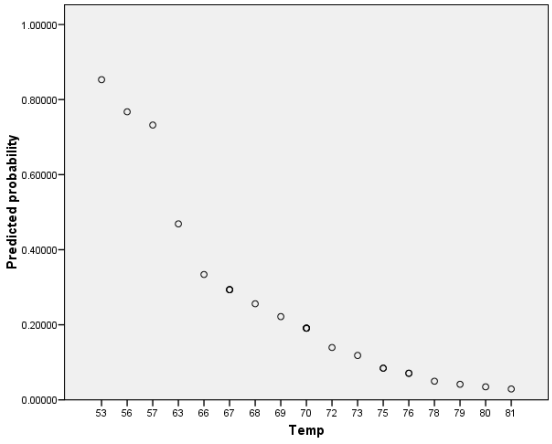
Let p= probability that the O-ring will fail. We will use SPSS to fit a logistic model to the data:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X + \epsilon$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Temp	-.188	.089	4.473	1	.034	.828	.696	.986
	Constant	11.746	6.021	3.806	1	.051	126299.6		

a. Variable(s) entered on step 1: Temp.



(a) Interpret the SPSS output. Write down the fitted model.

(b) Why does the scatterplot of the estimated probabilities against observed temperatures not look like a (nice) S-shaped logit curve?

(c) On the morning of January 28, 1986 the air temperature was about 31°F. Even though this value is outside the range of observed temperatures, use the logit model to predict the probability of O-ring failure for the Challenger flight.

Polynomial Regression

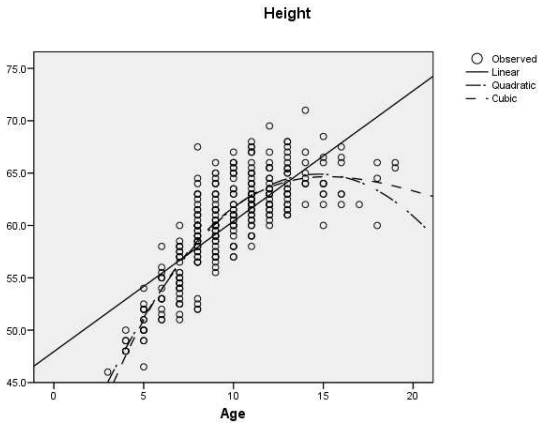
Definition: The k^{th}-DEGREE POLYNOMIAL REGRESSION MODEL EQUATION is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \epsilon$$

where $\epsilon \sim \text{Normal}(0, \sigma^2)$.

Similarly to the case of simple linear regression (where $k = 1$), the parameters $\beta_0, \beta_1, \ldots, \beta_k$, and σ^2 can be estimated by minimizing the sum of squared residuals. For example $\hat{\sigma}^2$ =MSE. SPSS will estimate all these model parameters for you, together with confidence intervals.

Example: A study on child-respiratory disease recorded (among many other things) the heights and ages of 318 girls aged 3-19. Below find a scatterplot of height against age and the SPSS results of fitting a linear, quadratic and cubic regression model.



Model Summary and Parameter Estimates

Dependent Variable: Height

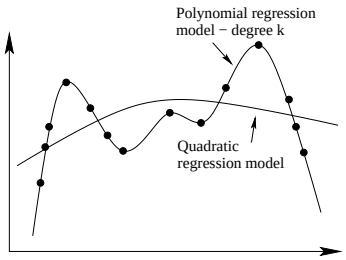
	Model Summary					Parameter Estimates			
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	.581	438.098	1	316	.000	47.953	1.245		
Quadratic	.716	397.121	2	315	.000	33.563	4.255	-.144	
Cubic	.719	267.527	3	314	.000	28.866	5.794	-.298	.005

The independent variable is Age.

(a) Which polynomial regression model do you think is most appropriate for this case?

(b) Write down the model equation (with numbers where possible).

Remark: If your data contains n observations, then a polynomial of degree $n-1$ can be found that achieves a perfect fit (all $\epsilon = 0$, $R^2 = 1$). Is this the "best" model? Probably not. A less complicated model (with a polynomial function of lower degree) would be preferable.



Recall that for the simple linear regression model the coefficient of determination was defined as

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

This same R^2 can also be computed for a nonlinear regression model. In general, raising the degree of the polynomial in the regression equation can only raise R^2 (more parameters in the model mean we can achieve a better fit). If the quantity R^2 is to be used to select the "best" model, then we need to build in a penalty for overly complicated models.

DEFINITION: The ADJUSTED COEFFICIENT OF DETERMINATION is defined as

$$\text{adjusted } R^2 = 1 - \frac{n-1}{n-(k+1)}\frac{\text{SSE}}{\text{SST}}$$

where n is the number of observations, and k is the degree of the polynomial model.

Other than R^2, the adjusted R^2 coefficient of determination can be used as a guide for model selection. The higher the adjusted R^2, the better the model. You still have to also check residual plots, to make sure that the model is actually appropriate.