Statistical Models

Statistical Regression Models

Índice

1. Statistical (Regression) Models
2. Multiple Linear Regression
3. Diagnostics and Validation
4. Predictions
5. Linearized Relationships
6. Nested Models

The role of statistical Models

All models are wrong, but some are useful (Box, 1965)

$$data = model + residual$$

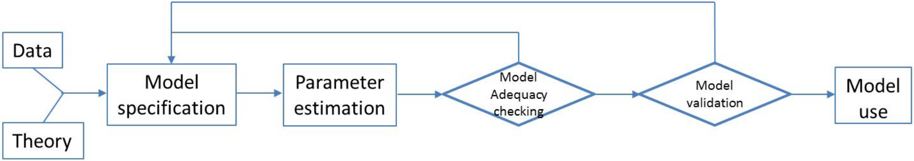
- The model is the underlying, simplified structure of set of data.
- The residual represents the difference between the model and the observed data points. The residual souldn't contain any additional pattern or structure. If it does contain additional structure, then the model associated with the data needs refinement and the process should continue until all parts of this structure can be assumed to be contained within the model.

The role of statistical Models

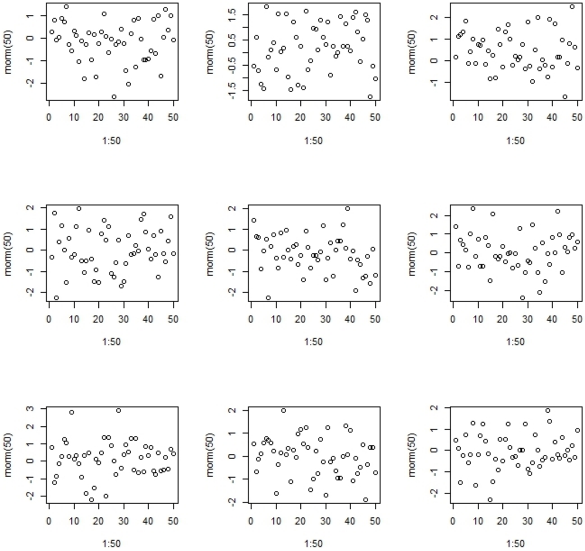
Everything should be made as simple as possible, but not simpler. (The principle of Occam's razor).

Data analysis is an art (subjective decisions!) based on science (objective tools!). We might therefore call data analysis an artful science! (Online Course Stat 501 at Penn State)

Regression model-building process

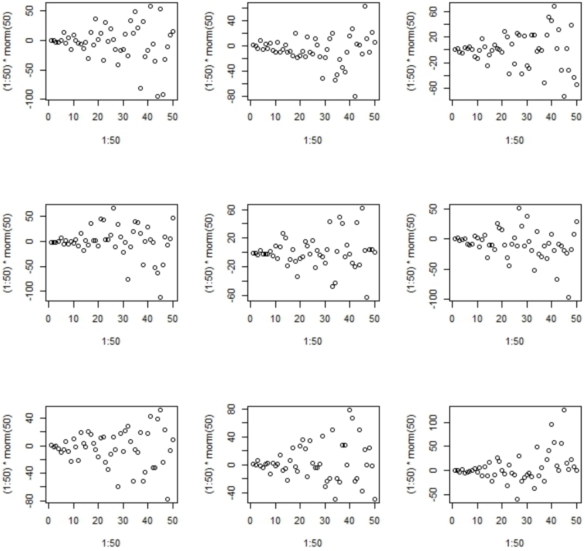


What residuals should look like:



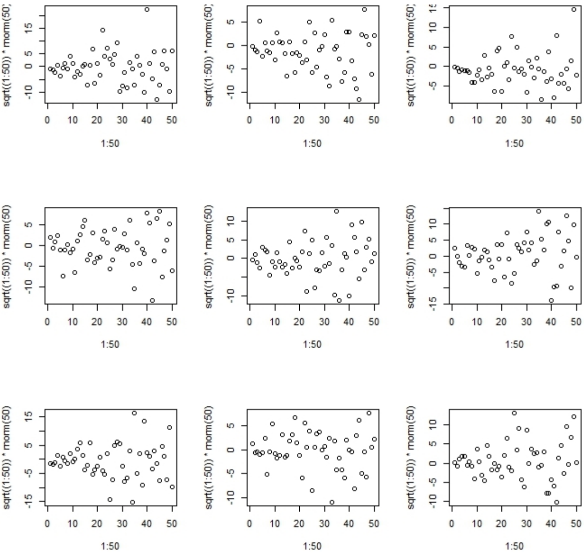
What residuals should not look like:

Severe Heterocedasticity (variance non constant)



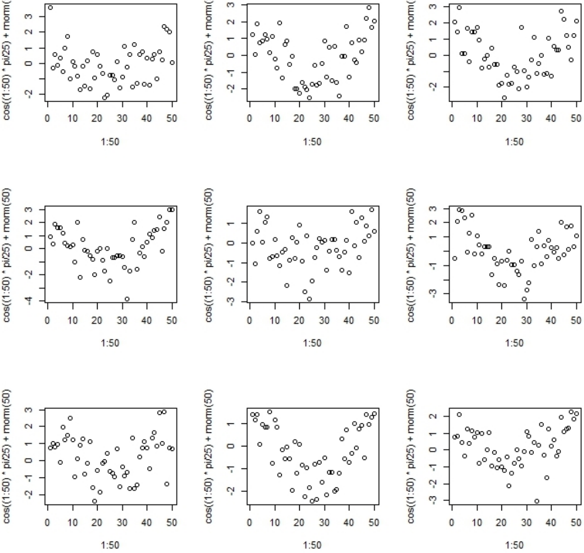
What residuals should not look like:

Mild Heterocedasticity (variance non constant)



What residuals should not look like:

Showing non independence



Data structure

Consider the data

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{x_1} = \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix} \quad \mathbf{x_2} = \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2n} \end{pmatrix} \quad \cdots \quad \mathbf{x_p} = \begin{pmatrix} x_{p1} \\ x_{p2} \\ \vdots \\ x_{pn} \end{pmatrix}$$

and the generic model

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon$$

or

$$E[\mathbf{y}|\mathbf{x}] = \mathbf{f}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p})$$

where $\mathbf{y}$ and ϵ are assumed to be random vectors and the matrix X contains n observations on variables $(X_1, \ldots, X_p)$.

General Linear Model

First order approximation, multiple linear model:

$$\underbrace{y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{Hyperplane}} + \epsilon$$

or

$$E[y|x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Assumptions of the linear model (later). They support inference and prediction.
- Estimation by Least Squares, Maximum Likelihood,...
- Interpretable, computation is fast.
- Simple interactions and transformations are easy.
- Dummy variables allow the use of categorical information.

Generalized Linear Model (1972)

$$g(E[y|x]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- For example: logistic regression model, Poisson regression model, ...
- Flexible: availability of alternative link functions.
- They relax some of the assumptions of the linear model.
- Estimation: Newton-Raphson algorithm, Iteratively-reweighted least squares (McCullagh and Nelder, 1989).

Additive Model (1985)

$$E[y|x] = \beta_0 + f_1(x_1) + \cdots + f_p(x_p)$$

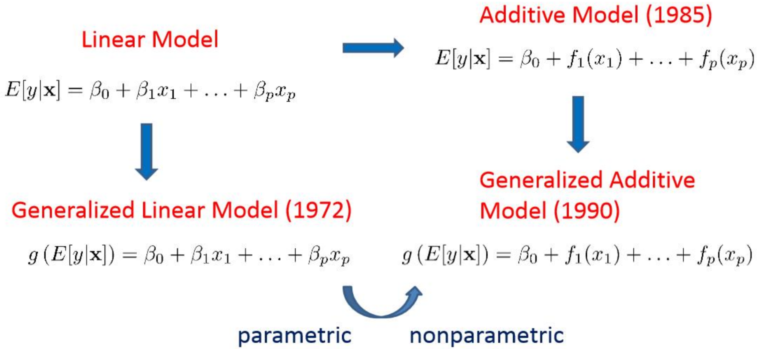
- Flexible, non parametric model.
- f_k are unknown smooth functions fit from the data.
- The parameters are $\{f_k\}, \beta_0$ and σ^2.
- Estimation: Backfitting Algorithm (Breiman and Friedman, 1985)

Generalized Additive Model (1990)

$$g(E[y|x]) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p)$$

- For example, an additive version of logistic regression.
- Flexible, non parametric model.
- f_k are unknown smooth functions fit from the data.
- The link function is chosen by the user based on domain knowledge.
- Estimation: IRLS + Backfitting (Hastie and Tibshirani, 1990).

Models in a nutshell



Linear Model

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Additive Model (1985)

$$E[y|\mathbf{x}] = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

Generalized Linear Model (1972)

$$g\left(E[y|\mathbf{x}]\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Generalized Additive Model (1990)

$$g\left(E[y|\mathbf{x}]\right) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

parametric nonparametric

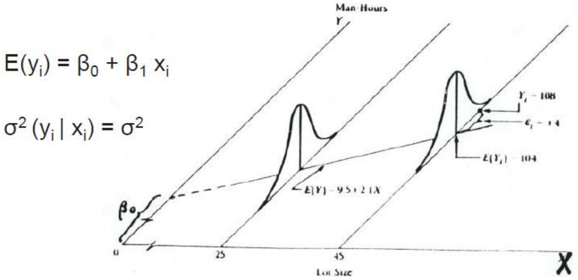
©Wang, Jiang 2013

Simple Linear Regression Model

In its simpler form, the linear model becomes:

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{fixed}} + \underbrace{e}_{\text{random}}, \quad e \sim IIDN(0, \sigma^2)$$

For each x there is a (hypothetical) distribution of y values with means linearly related to x and constant variance.



$E(y_i) = \beta_0 + \beta_1 x_i$

$\sigma^2 (y_i \mid x_i) = \sigma^2$

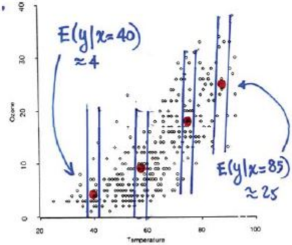
Simple Linear Regression Model

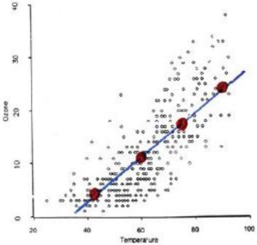
How does atmospheric ozone depend on temperature?

In general, we want to be able to describe / predict how a response *y* is related to one (or more) explanatory variables (x).

Such a description is always approximate.

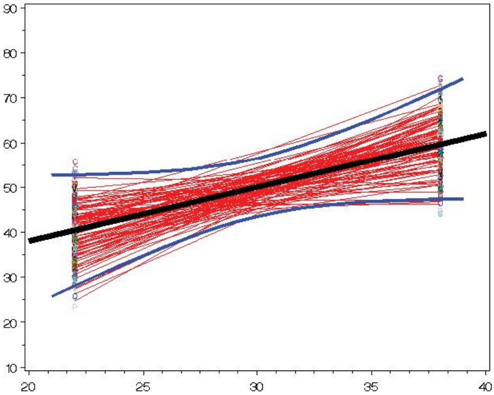
Nevertheless, the model can be extended.





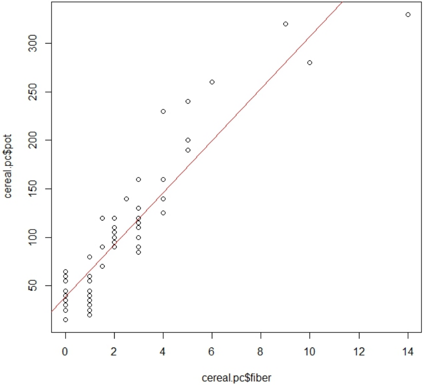
Simple Linear Regression Model

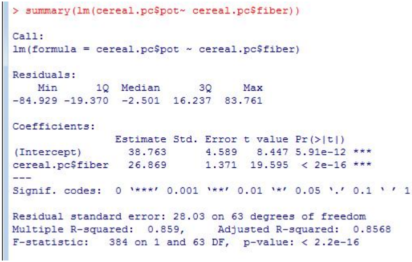
We observe a sample of data and fit one linear regression line. Using sample estimates, their properties and sampling distributions, we draw conclusions about the population values, we perform statistical test, etc.:



Regression lines for 500 samples.

Simple Linear Regression Model



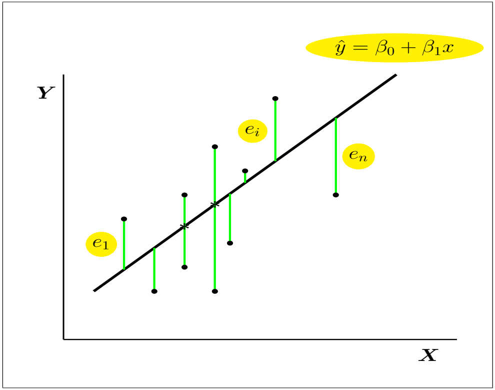


```
> summary(lm(cereal.pc$pot~ cereal.pc$fiber))

Call:
lm(formula = cereal.pc$pot ~ cereal.pc$fiber)

Residuals:
    Min      1Q  Median      3Q     Max 
-84.929 -19.370  -2.501  16.237  83.761 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       38.763      4.589   8.447 5.91e-12 ***
cereal.pc$fiber   26.869      1.371  19.595  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.03 on 63 degrees of freedom
Multiple R-squared:  0.859,     Adjusted R-squared:  0.8568 
F-statistic:   384 on 1 and 63 DF,  p-value: < 2.2e-16
```

Simple Linear Regression Model

Least squares criterion implies minimizing:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2$$



Simple Linear Regression Model

We obtain estimates for the model parameters: $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma^2}$.

Parameter interpretation:

- $\hat{\beta}_0$, the intercept, is the value of the response for $X = 0$. Most of the time it has not a meaning, or sensible or realistic interpretation itself.
- $\hat{\beta}_1$, the slope, represents the variation in the response Y when X is increased (or decreased) one unit.

Simple Linear Regression Model

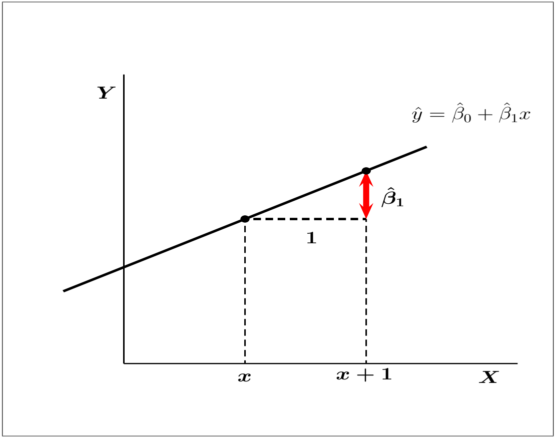
$\hat{\beta}_1$ represents the variation in the response Y when X is increased (or decreased) one unit.

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$$\hat{y}(x+1) = \hat{\beta}_0 + \hat{\beta}_1 (x+1),$$

Thus,

$$\hat{y}(x+1) - \hat{y}(x) = \hat{\beta}_1.$$



Simple Linear Regression Model

Least squares estimation doesn't provide any information about the variability of the experimental error. This information is in the residuals.

That's where comes into play the (famous) assumptions on the Linear Regression Model. Normality is one of them. This leads to another estimation method, Maximum Likelihood Estimation. The estimates are pretty much the same as those obtained with the Least Squares method but their distributional properties permit to make inferences about the model parameters, predictions, etc.

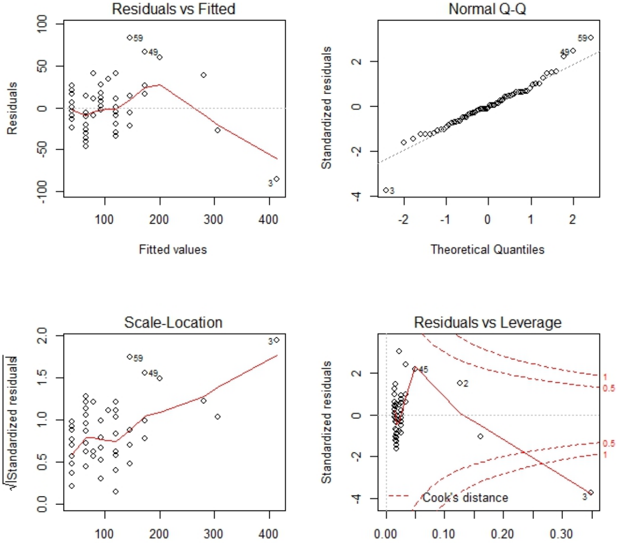
Simple Linear Regression Model

To use the model for inference and prediction, we need to impose four conditions that comprise the simple linear regression model (LINE). They can be stated in terms of the response variable Y_i or in terms of the residuals (errors) e_i:

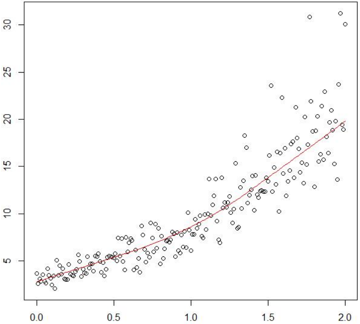
- Linearity: the mean of the response $E(Y_i|X_i)$ at each value of the predictor is a linear function of the X_i.
- Independence: the residuals (errors) e_i are independent.
- Normality: the residuals, for each value of the predictor X_i are normally distribuited.
- Equal variances (σ^2): the residuals, at each value of the predictor X_i have equal variances.

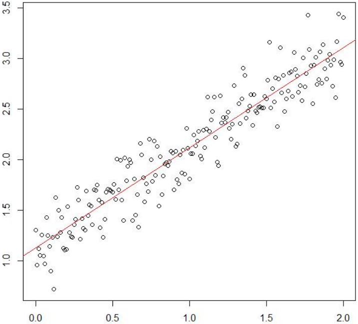
Simple Linear Regression Model

Diagnostics: Residual Plots.



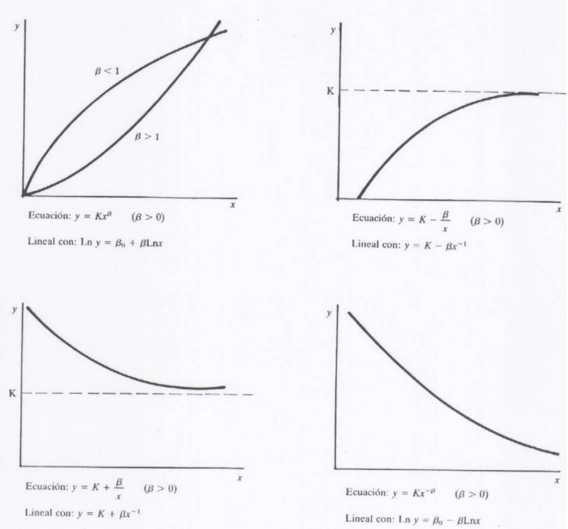
Simple Linear Regression Model Flexibility





Transforming one variable (or both) can turn a relationship into a linear one

Simple Linear Regression Model Flexibility



Ecuación: $y = Kx^{\beta}$ $(\beta > 0)$

Lineal con: Ln $y = \beta_0 + \beta$Lnx

Ecuación: $y = K - \frac{\beta}{x}$ $(\beta > 0)$

Lineal con: $y = K - \beta x^{-1}$

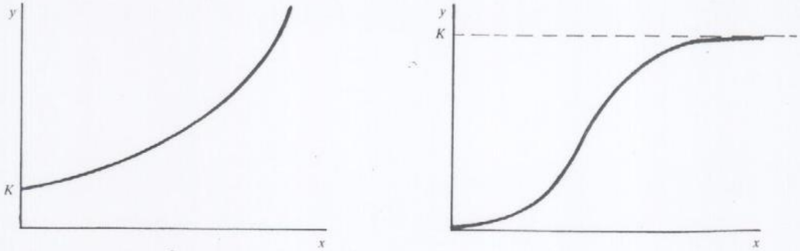
Ecuación: $y = K + \frac{\beta}{x}$ $(\beta > 0)$

Lineal con: $y = K + \beta x^{-1}$

Ecuación: $y = Kx^{-\beta}$ $(\beta > 0)$

Lineal con: Ln $y = \beta_0 - \beta$Lnx

Simple Linear Regression Model Flexibility



Ecuación: $y = Ke^{\beta x}$ $(\beta > 0)$

Lineal con: Ln $y = \beta_0 + \beta x$

Ecuación: $y = Ke^{-\beta/x}$

Lineal con: Ln $y = \beta_0 - \beta x^{-1}$

Multiple Linear Regression Model

Multiple regression linear model has the form:

$$\underbrace{y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{Hyperplane}} + e$$

- There are multiple predictors (independent vars.) of varying kinds and a single outcome (dependent variable).
- Goals: prediction vs. explanation, not mutually exclusive.

Statistical Assumptions

- Normality (and Linearity): the conditional distribution of Y for any combinations of values of the $X_1, \ldots, X_p$ is Normal. The expected value is a linear function of the $X's$,

$$(Y|X_1 = x_{i1}, \ldots, X_k = x_{ik}) \approx N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \sigma^2)$$

$$e_i \approx N\left(0, \sigma^2\right)$$

- Independence. The observations from Y are statistically independent.
- Homocedasticity. The conditional variance of Y given any specific combination of values of the $X_1, \ldots, X_p$ is the same, i.e., σ^2.

Challenges

- *Choosing the best model:* which subset of variables perform well?
- *Interactions between variables:* in some cases, independent variables interact, and the regression equation will not be accurate unless this interaction is taken into account.
- *Difficulties visualizing the regression relationships:* with two predictors, there is a regression surface and, with 3 predictors or more, you run out of dimensions for plotting.
- *Model interpretation becomes substantially more difficult than when you have only one predictor variable.*

Coefficient Interpretation

Each β_j is a partial regression coefficient, reflecting the change in the independent variable per unit change in the $j-$th independent variable, *assuming all other independent vars. are held constant*. The coefficients of a multiple regression must not be interpreted marginally!

So a regression coeffcient in multiple regression is a partial regression coefficient between a variable and the criterion, with all the other regressors partialled out of both variables.

Regression Tests

After fitting a model to a set of data it is necessary to asses the adequacy of the fit. From the general assumptions made on the error terms (and on the distribution of Y) different tests are developed.

- Overall regression test (F test):

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus:

$$H_1 : \text{There exists some } \beta_i,\ i = 1, \ldots, k \text{ so that } \ \beta_i \neq 0.$$

p-value < 0.05 will reject H_0, for instance:

F-statistic: 694 on 3 and 140 DF, p-value: $<$ 2.2e-16

Regression Tests

- Failing to reject H_0,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

  may be due to:
  - There is no linear relationship between the response variable and the explanatory variables
- Rejecting H_0 means the model with some (or all) predictors is better to explain the output Y than a model with just the mean of the y scores.
  - More specific tests are available for identifying important explanatory variables (t-tests).

t tests (each variable’s coefficient at a time)

```
> summary(modelo3)

Call:
lm(formula = X2 ~ X3 + X4 + X5, data = diabetis, x = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max 
-39.223 -10.646   0.927   8.559  71.432 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 26.172321   4.135995   6.328 3.14e-09 ***
X3           0.193768   0.007941  24.400  < 2e-16 ***
X4          -0.039134   0.013483  -2.902   0.0043 ** 
X5          -0.013996   0.022427  -0.624   0.5336    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.28 on 140 degrees of freedom
Multiple R-squared: 0.937,      Adjusted R-squared: 0.9356 
F-statistic:   694 on 3 and 140 DF,  p-value: < 2.2e-16
```

The intercept and coefficients for vars. $X3$ and $X4$ may be considered different from 0. Variable $X5$ may be dropped out, given the others in the model.

Relation between t tests and overall F test

Scenario	F Test	specific tests
1	Significant	Every one Significant
2	Significant	Some of them Significant
3	Significant	None Significant
4	Not Significant	All of them Significant
5	Not Significant	Some of them Significant
6	Not Significant	None Significant

Posible scenarios

Scenarios 1 and 2: proceed with the analysis, dropping out variables if needed.

Scenarios 3, 4 and 5: Multicollinearity problems.

Scenario 6: No linear relationships are detected between the variables involved in the model.

Coefficient of Determination

Once the multiple regression model is fit,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k,$$

we can add new columns to our data set $\hat{Y}$, and e_i:

	X_1	$\cdots$	X_k	Y	$\hat{Y}$	e_i
Observ. 1	x_{11}	$\cdots$	x_{k1}	y_1	$\hat{y}_1$	$y_1 - \hat{y}_1$
Observ. 2	x_{12}	$\cdots$	x_{k2}	y_2	$\hat{y}_2$	$y_2 - \hat{y}_2$
	$\vdots$				$\vdots$	
Observ. n	x_{1n}	$\cdots$	x_{kn}	y_n	$\hat{y}_n$	$y_n - \hat{y}_n$

Where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$.

Coefficient of Determination

- The correlation between the predicted scores ($\hat{Y}$) and the observed scores (Y) is called the multiple correlation coefficient (R).
- R^2 measures the proportion of the variance of the dependent variable about its mean that is explained by the independent, or predictor, variables. Low values may be explained because important variables have been left out of the model.
- The addition of independent variables will always cause R^2 to rise. A modified measure, adjusted R^2, takes into account the nuber of independent variables included and the sample size.

$$R_a^2 = R^2 - \frac{k(1 - R^2)}{n - k - 1}$$

k is the number of predictors. It can be used to compare models with different number of predictors but with the same Y.

Diagnostics and Validation

Regression model building is an iterative process. The first model we try may prove to be inadequate. Regressions diagnostics are used to detect problems:

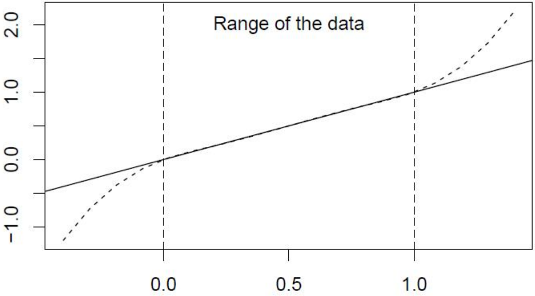
- Residuals that aren't:
  1. Normal distributed
  2. Homocedastic (non constant variance)
  3. Independent
- Non linearity, model misspecification
- Multicollinearity

Problems should be detected and improvements sould be suggested. This is a hands-on process.

PREDICTIONS

They should be done within the data range. The new x should be within the range of validity of the model. If not, prediction may be quite unrealistic. There are many risks at extrapolating:

If she loves you more each and every day,
by linear regression she hated you before you met.



model in solid, real relationship dotted line

Linearized Relationships

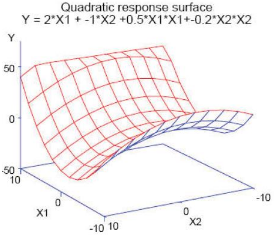
The model is linear *in the parameters*. Its versatility permits to consider models like:

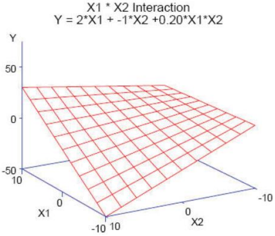
- Quadratic polynomial regression model, one predictor:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

- Quadratic polynomial regression model, two predictors and interaction:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + e$$





Linearization of relationships

- $y = \beta_0 + \beta_1 \left(\frac{1}{x_1}\right) + \beta_2 \ln x_2 + \beta_3 \sqrt{x_3} + e$
- Model with interactions:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

- Multiplicative Model:

$$y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} e,$$

  where e is the random error.

  - Taking logarithms:

$$\ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \delta \ln x_3 + lne,$$

  which is a linear model:

$$Y = \beta_0 + \beta z_1 + \gamma z_2 + \delta z_3 + e$$

Non linear relationships

However, these relationships can't be linearized:

- $$y = \beta_0 + \beta_1 e^{-\beta_2 X} + e$$
- $$y = \beta_0 + \beta_1 X + \beta_2 (\beta_3)^X + e$$
- For instance:
  $$y = \beta_0 + \beta_1 2^{-\frac{x}{\beta_3}} + e$$
  is linear in the parameters β_0 and β_1 but non linear in parameter β_3. Using Nonlinear Least Squares (for instance, function **nls()** in R) requires specifying the form of the model and having starting values for the iterative estimation of the parameters and this can be difficult to determine in general cases.

General Linear Model

- ANOVA as Dummy Variable Regression: Suppose we have 3 groups, and we want to test the null hypothesis that all 3 come from populations with the same mean. A side assumption is that all groups have the same variance, and that the population distributions are normal. The alternative hypothesis is that at least one of the groups has a mean that is different from the others. Linear Regression Model with factor variables.
- ANCOVA as regression with categorial variables and one continuous variable X.

Nested Models

- One model is nested within another if it is a special case of the other in which some model coeffcients are constrained to be zero.
- When two models are nested multiple regression models, there is a simple procedure for comparing them.
- This procedure tests whether the more complex model is significantly better than the simpler model.
- In the sample, of course, the more complex of two nested models will always fit at least as well as the less complex model.
- This is done via Partial F-tests.

Nested Models. Example.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e \quad mod1$$

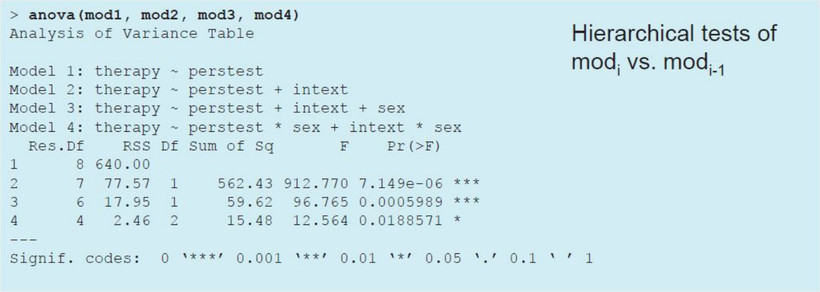
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 e \quad mod2$$

R will perform the partial F-test automatically, using the anova command. We are testing $H_0 : \beta_4 = \beta_5 = 0$.

```
> anova(mod1, mod2)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X1:X2
Model 2: Y ~ X1 + X2 + X1:X2 + I(X1^2) + I(X2^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     94 29.422
2     93 29.420  1 0.0022638 0.0072 0.9328
```

With such a p-value we prefer the simpler model (mod 1).

Nested Models. Example.



```
> anova(mod1, mod2, mod3, mod4)
Analysis of Variance Table

Model 1: therapy ~ perstest
Model 2: therapy ~ perstest + intext
Model 3: therapy ~ perstest + intext + sex
Model 4: therapy ~ perstest * sex + intext * sex
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1      8 640.00
2      7  77.57  1    562.43 912.770 7.149e-06 ***
3      6  17.95  1     59.62  96.765 0.0005989 ***
4      4   2.46  2     15.48  12.564 0.0188571 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hierarchical tests of mod$_i$ vs. mod$_{i-1}$