1 Multiple Regression Models

In this section the Multiple Regression model will be introduced as a mean of relating one numerical response variable y to two (or more) independent or predictor variables.
We will consider first order and interaction models and discuss the implications of choosing the different models.
Specifically we will consider the case that we have one numerical and one categorical predictor variable to model a third response variable. For example modelling the mean prestige of an occupation (response) in dependency on income (numerical predictor) and the type of occupation (blue collar, white collar, and professional) (categorical predictor).
How to use sample data to obtain estimates and confidence intervals for the model parameters and how to interpret these will be presented.

The Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

where

- y is the dependent variable
- $x_1, x_2, \ldots, x_k$ are the independent variables
- $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ is the deterministic part of the model
- β_i determines the contribution of the independent variable x_i
- e is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ.

We will now discuss different choices for the independent variables, then when building a model for a certain problem, we can choose out of those discussed.

1.1 First-Order-Model

The term *first* indicates that the independent variables are only included in the first power, we later see how we can increase the order.

The First-Order Model in Numerical Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

where $x_1, x_2, \ldots, x_k$ are quantitative variables, which are not functions of each other.
β_0 is the mean of y, when all independent variables equal 0.
β_i is the slope of the line relating y with x_i when all other independent variables are held fixed.

When choosing estimates for the parameters of the model, the criteria is the same as for the simple linear regression model
Make

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

as small as possible by choosing $b_1, \ldots, b_k$, where $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$.
Finding the estimates involves solving $k+1$ equations for $k+1$ parameters. This is *easy* with methods from linear algebra, but tedious without.
Here we will rely on the computer to find the estimates for us. The estimate for σ will be still the square root of the MSE.

$$\hat{\sigma} = s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-(k+1)}}$$

The degrees of freedom for Error are $df = n - (k+1)$.

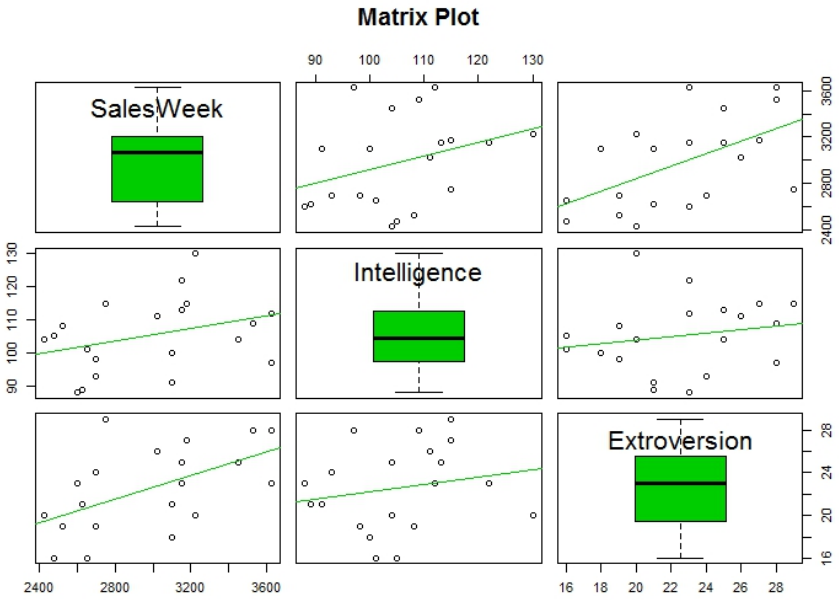
Example 1
(Adapted from http://simon.cs.vt.edu/SoSci/converted/MRegression/)
The ABC corporation decides to study the sales staff at their stores to determine if intelligence and extroversion (i.e., a friendly and outgoing personality) predict sales performance of current employees.
To conduct the study, all retail sales employees at existing stores take psychological tests designed to measure intelligence and extroversion. Also, past sales performance data is checked for each employee. In the end, there are three scores for each of 20 sales people:

1. an intelligence score (on a scale of 50-low intelligence to 150-high intelligence),
2. an extroversion score (on a scale of 15-low extroversion to 30-high extroversion), and
3. sales performance expressed as the average dollar amount sold per week.

We propose the following model:

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

y = sales, x_1 = intelligence, x_2 = extroversion, e = error (deviation from regression function), with $e \sim \mathcal{N}(0, \sigma)$.
In a first step the predictors should be plotted against each other and against the the response variable to check if the model seems reasonable.



The multiple linear regression model might be a reasonable description of the relationship:
The plots of intelligence with sales performance, and extroversion with sales performance both show a weak positive linear relationship (maybe).
The scatter plot of the two predictors, intelligence and extroversion, does not show a linear relationship, which is what we hope for.

If the predictor variables share a strong linear relationship, one of them becomes redundant. For example variables one giving the temperature measured in Fahrenheit and the other temperature measured in Celsius would provide identical information, and only one of them should be included in the model. – Mathematically we create estimation problems when we include two highly related variables.

In a next step we should do a model utility test (what does that mean here?) Will hold the thought, and before look at the estimates for our parameters, we get from the data.

The R output for the analysis from this data is:

```
  Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   993.925    788.099   1.261   0.2243
Intelligence    8.220      7.013   1.172   0.2573
Extroversion   49.709     19.634   2.532   0.0215
---

Residual standard error: 332.1 on 17 degrees of freedom
Multiple R-squared:  0.3526,Adjusted R-squared:  0.2765
F-statistic:  4.63 on 2 and 17 DF,  p-value: 0.02482
```

Giving the estimated regression function

$$\hat{y} = 993.9 + 8.22(\text{Intelligence}) + 49.709(\text{Extroversion})$$

Interpretation:
The mean sales for sales people with intelligence = 0 and extroversion = 0 is estimated to be \$ 993.9 per week (meaningful?).
For every extra point on the intelligence scale the sales per week increase on average by \$ 8.22 per week when correcting for the influence of extroversion.
For every extra point on the extroversion scale the sales per week increase on average by \$ 49.71 per week when correcting for the influence of intelligence.

The estimate for σ equals $\sqrt{332.1} = 18.22$, which means, that on average the observed sales per week fall \$ 18.22 away from the estimated values from the regression function.

But is this model an appropriate description of the relationship between sales per week and intelligence and extroversion of the sales person? The scatter plots seem to say so but we better test.

Example 2
Feeding snow goslings

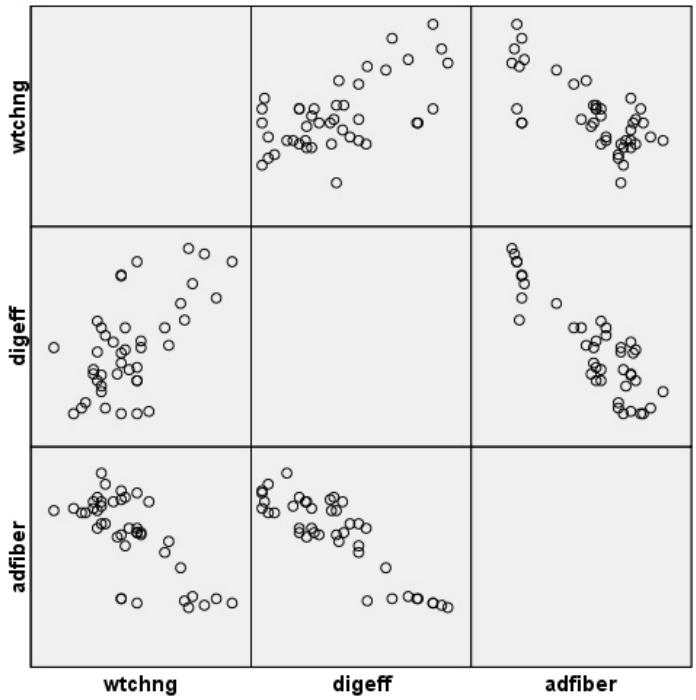
Botanists were interested in predicting the weight change of snow geese goslings as a function of digestion efficiency, acid-detergent fibre (amount in the food) (all measured in percentage), and diet (plants or duck chow (duck food)). The acid-detergent fibre is the least digestible portion in plant food.
Our goal will be to fit the first order model to the data, where we want to explain the weight change,

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

y = weight change, x_1 = digestion efficiency, x_2 = acid-detergent fibre, e=error.
In order to assure that a linear model is reasonable first obtain scattergrams for

1. the response variable with each explanatory variable, and
2. each pair of explanatory variables.

We hope to find that the response variable shows a linear relationship with the explanatory variables. At least we do not want to see a non linear relationship.
When analyzing the scattergrams for the pairs of explanatory variables, we do not want to find a strong linear relationship. Consider they would relate in a perfect linear relationship, then in the Multiple Linear Regression Model they would bear the information on the response variable and would be redundant.



The matrix plot shows that the relation ship between weight change and the other two variables show a linear trend, the linear model seem to be appropriate for describing the relationship between the variables.
But the scattergram for digestion efficiency and fibre displays a strong negative linear relationship, we must fear that one of them will show to be redundant and we will have to check for this effect.

The estimates of the coefficients and the confidence intervals for those produced by SPSS are:

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
(Constant)	12.180	4.402		2.767	.009	3.276	21.085
digeff	-.027	.053	.115	.496	.623	-.135	.082
adfiber	-.458	.128	-.826	-3.569	.001	-.717	-.198

a Dependent Variable: wtchng

From the table we get:

$$b_0 = 12.18, b_1 = -0.027, \text{ and } b_2 = -0.458$$

In the ANOVA table we find $MSE = 12.387$, which results in $s = \sqrt{12.387} = 3.5195$.

Interpretation: $\beta_1 = -0.27\%$ gives the mean change in the weight difference for a one percent increase in the digestion efficiency when keeping the amount of fibre being fed to the gosling constant. Based on the sample we estimate that the weight drops in average by 0.027 percent if the digestion efficiency increases one percent and the acid detergent fibre remains the same.
Caution: We estimate a drop in weight change when increasing the digestion efficiency, this is in contradiction to what we observe in the scattergram for weight change and digestion efficiency. The reason for the discrepancy between the two is the high correlation between digestion efficiency and the fibre variable, we already observed in the scattergram.
Try the interpretation of β_2 is the value consistent with the scattergram?
β_0 does not have a meaningful interpretation in this example, it would be the mean change in weight if the digestion efficiency is at 0 percent and the acid-detergent fibre is at 0 %.

1.2 The Overall Model Utility and Inference about the Parameters

The multiple coefficient of determination, R^2, is calculated by

$$R^2 = \frac{\text{Explained variability in } y}{\text{Total variability in} y}$$

Just as in the simple linear regression model the coefficient of determination measures how much of the sample variability in y can be explained through the linear model applied.

Example 3
For the sales example the out put shows $R^2 = 0.3526$, about 35% of the variation in the sales per week can be explained of the differences in intelligence and extroversion.– Not that much!

For the gosling example the output shows an $R^2 = 0.529$. About 53% of the variation in the weight change can be explained through different values in acid-detergent fibre and digestion efficiency.

R^2 close to one indicates a very good fit of the model, and R^2 close to zero indicates lack of fit. One needs to be careful though, because if the sample size equals the number of parameters $R^2 = 1$, and a high number of parameters ensures a goof fit, even though the population might

not be represented by the estimated model. In order to avoid this pitfall one should use a much larger sample than parameters when fitting a linear regression model.

Instead of just interpreting R^2, one can test for model utility.
If all slope parameters would be 0, then nothing about the response variable could be learned from the predictor variables in the model. In a first step, we will try to disprove that all slopes in the model are all zero and test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$ versus H_a : at least one slope is not 0.
The test statistic is based on

$$F = \frac{(SS_{yy} - SSE)/k}{SSE/(n-k-1)}$$

which is the ratio of the (standardized) explained variability to the (standardized) unexplained variability.
A large ratio indicates that most of the variability could be explained, and a small value indicates that barely any variability could be explained. To judge if the ratio is small or large we have to take the degrees of freedom into account.

Model-Utility Test: F-Test for overall usefulness of the model

1. Hypotheses: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ versus H_a : at least one is not zero. Choose α.
2. Assumptions: The assumption of the multiple regression model hold up.
3. Test Statistic:
$$F_0 = \frac{(SS_{yy} - SSE)/k}{SSE/(n-k-1)} \text{ with } df_n = k, df_d = n-k-1.$$
4. P-value = $P(F > F_0)$ (this is what we will focus on).
5. Reject H_0 if P-value$< \alpha$, other wise do not reject H_0.
6. Put into context.

Continue Sales example:

1. Hypotheses: $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : at least one is not zero. Choose $\alpha = 0.05$.
2. Assumptions: ...
3. Test Statistic from output:

```
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   993.925    788.099   1.261   0.2243
Intelligence    8.220      7.013   1.172   0.2573
Extroversion   49.709     19.634   2.532   0.0215
---

Residual standard error: 332.1 on 17 degrees of freedom
Multiple R-squared:  0.3526,Adjusted R-squared:  0.2765
F-statistic:  4.63 on 2 and 17 DF,  p-value: 0.02482
```

$$F_0 = 4.63, df_n = 2, df_d = 17.$$

4. P-value = $P(F > F_0)$=0.02482
5. Reject H_0 since P-value< $\alpha = 0.05$
6. At significance level of 0.05 the data provide sufficient evidence that that the model is useful in explaining the sales per week by considering intelligence and extroversion.

Continue Gosling example:

1. Hypotheses: $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : at least one is not zero. Choose α.
2. Assumptions: The assumption of the multiple regression model hold up, we checked that by looking at the scattergram that the model seems to fit reasonably well.
3. Test Statistic from output:

ANOVA(b)

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	542.035	2	271.017	21.880	.000(a)
Residual	483.084	39	12.387		
Total	1025.119	41			

a Predictors: (Constant), adfiber, digeff

b Dependent Variable: wtchng

$$F_0 = \frac{(SS_{yy} - SSE)/k}{SSE/(n-k-1)} = 21.880$$

with $df_n = 2, df_d = 42 - 2 - 1 = 39$.

4. P-value = $P(F > F_0)$=.000
5. Reject H_0 since P-value< $\alpha = 0.05$
6. At significance level of 0.05 the test confirms our conclusion based on the coefficient of determination, we have sufficient evidence that the model is useful in explaining the weight change of snow geese goslings.

Once we established that the model seems to describe the relationship between the response and the independent variables properly, we start asking questions about the values of the parameters, and which of them show a significant influence on the response variable.
We can answer this question in two ways, we can find confidence intervals for the parameters, or conduct tests.

A $C \times 100\%$ Confidence Interval for a slope parameter β_i

$$\hat{\beta}_i \pm t^* s_{b_i}$$

The estimates $\hat{\beta}_i$ and the standard error $s_{\hat{\beta}_i}$, should be obtained through SPSS, formulas are too complex to be presented here.

Continue Sales example:
From R: 95% Confidence interval for β_1 (slope for Intelligence): [-6.575296, 23.01512]
95% Confidence interval for β_2 (slope for Extroversion): [8.285059, 91.13221]
Since 0 falls with in the 95% CI for β_1, we conclude that β_1 might be 0, and that intelligence has no significant influence on sales peer week of a sales person.
Since the confidence interval for β_2 does not include 0, we are 95% confident that $\beta_2 \neq 0$ and $\beta_2 > 0$, that is the more extroverted a sales person the higher the sales per week.

Continue Gosling example:
From SPSS: Confidence interval for β_1: [-0.135, 0.082]
Confidence interval for β_2: [-0.717, -0.198]
Since 0 falls with in the 95% CI for β_1, we conclude that β_1 might be 0, and the digestion efficiency has no significant influence on the weight change.
Since the confidence interval for β_2 does not include 0, we are 95% confident that $\beta_2 \neq 0$ and $\beta_2 < 0$, that is the less acid-detergent fibre in the food, the more the goslings gain weight.
Another way of answering this question is the **test of an individual slope parameter β_i.**

1. Hypotheses: Parameter of interest β_i.

test type	hypotheses
upper tail	$H_0 : \beta_i \leq 0$ vs. $H_a : \beta_i > 0$
lower tail	$H_0 : \beta_i \geq 0$ vs. $H_a : \beta_i < 0$
two tail	$H_0 : \beta_i = 0$ vs. $H_a : \beta_i \neq 0$

   Choose α.

2. Assumptions: Regression model is appropriate

3. Test statistic:
$$t_0 = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}, \quad df = n - k - 1$$

4. P-value:

test type	P-value
upper tail	$P(t > t_0)$
lower tail	$P(t < t_0)$
two tail	$2P(t > abs(t_0))$

5. Decision: If $P - value < \alpha$ reject H_0, otherwise do not reject H_0.

6. Put into context

Continue example: Let us test

1. Hypotheses: Parameter of interest β_2.
   lower tail $H_0 : \beta_2 \geq 0$ vs. $H_a : \beta_2 < 0$

   Choose $\alpha = 0.05$.

2. Assumptions: Regression model is appropriate, checked before.

3. Test statistic:
$$t_0 = -3.569, \quad df = 39$$

4. P-value:

   The two tailed is 0.001, since we are doing a lower tailed test and the test statistic is negative, the P-value=0.001/2=0.0005

5. Decision: Since $P - value < \alpha$ reject H_0.

6. The test gives the same result as the confidence interval, we found significant evidence that increasing the acid-detergent fibre in the food results in a drop in the weight change.

Be careful interpreting the result, when failing to reject $H_0 : \beta_i = 0$. This could be due to

1. there is no relationship between y and x_i, or

2. there is a relationship, but it is not linear, or

3. a type II error occurred.

1.3 Residual Analysis – here for information only

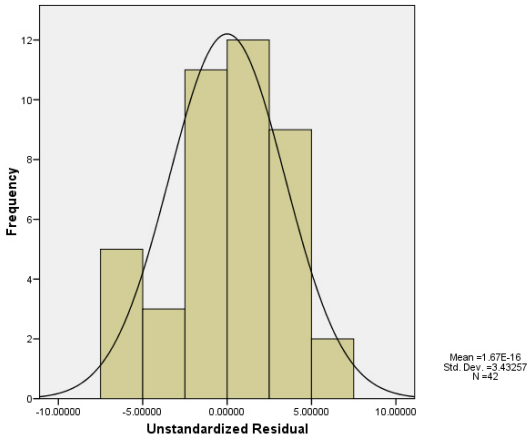
Before we can apply the model for predicting and estimating we should check if the assumption in the model are met.
We have to assume that the error, e, in the regression model is normally distributed with mean 0 and standard deviation σ. In order to check these assumptions the residuals should be obtained and

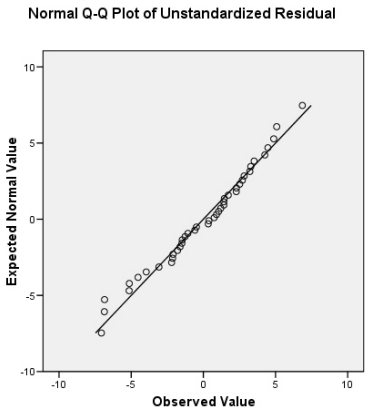
1. be displayed in a histogram and/or QQ-Plot, to asses if the assumption that they come from a normal distribution is reasonable, and

2. residual plots for all predictor variables (scattergram of predictor and residuals) should be obtained to illustrate that the standard deviation of the error does not depend on the value of the predictors.

Continue Gosling example:

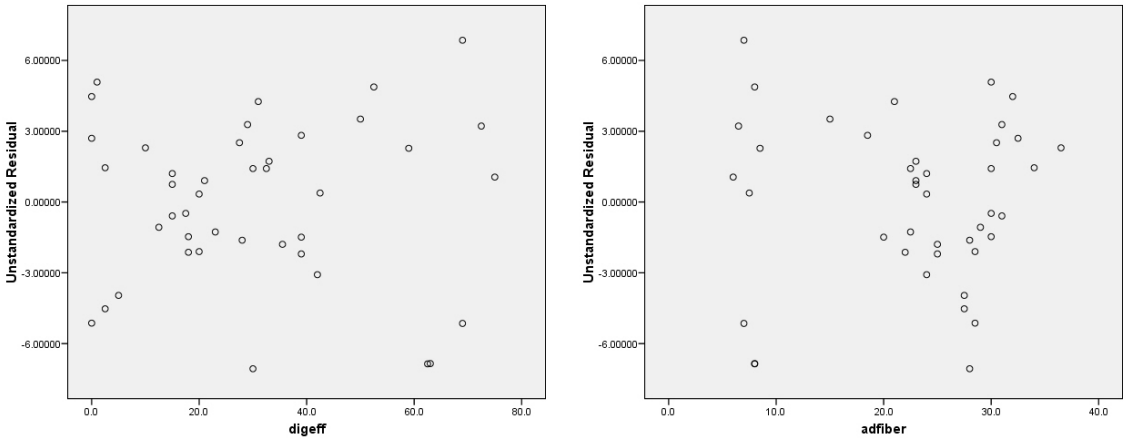
1. The histogram and QQ-plot for the residuals in our example:





Both graphs do not indicate a strong deviation from normality. They are no cause for concerns, about the assumptions in the model not being met.

2. In addition we need to check that the values of the predictor variables do not influence the spread:



Both graphs display even scatter around zero, which is not changing for different values of "digeff", and "adfibre".
The Residual Analysis support that the assumptions of the model are met, and for that reason our conclusion derived so far are valid.

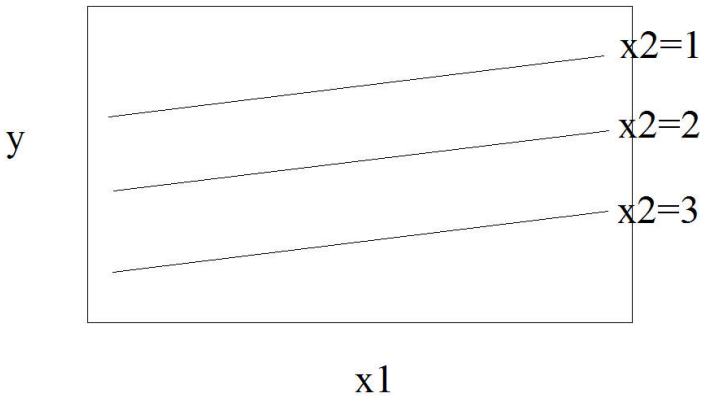
Now that we established that the model is appropriate to describe the relationship between y and the predictor variables and know which of the variables significantly influences the response variable, we can use the model to predict future values of y and estimate $E(y)$ for certain values of the predictor variables.

1.4 Interaction Models

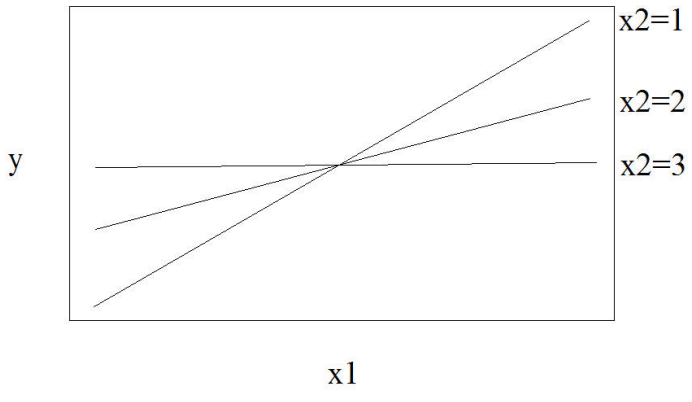
The First order model implies, that the response variable depends on 2 or more predictor variables in a linear fashion. For simplicity assume only 2 predictors.
To investigate the relationship between x_1 and y, assume we know $x_2 = x_2^*$ (for some fixed value x_2^*), then

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* = (\beta_0 + \beta_2 x_2^*) + \beta_1 x_1$$

which is a line in x_1 with slope β_1 and intercept $\beta_0 + \beta_2 x_2^*$ (this part does not depend on x_1).
In conclusion: When we check how μ_y depends on x_1 we observe that for all possible values of x_2 the slope in the relationship between x_1 and y is the same.



The lines for different values of x_2 are parallel (same slope, different intercepts)!
This means, when we consider a first order MLRM we inply that the relationship between each independent variable and the response is not affected by the value of the other independent variables. (In our example: the effect of intelligence on the sales is the same for all score on extroversion – is this reasonable?)
In many cases this seem quite unreasonable and in order to use a model that better describes the population one should include the possibility of interaction.
Two explanatory variables interact in their effect on the response, if the relationship between one of the explanatory variables and the reponse depends on the value of the second explanatory variable.
E.g. the relationship between height and weight is different for women and men, we say, Sex and hight interact in their effect on weight.
We will now include an interaction term with the model that will permit to check if this assumption is true. We will permit that the slopes of the lines relating x_1 with y differ for different values of x_2, we call this interaction in the regression model.



A Multiple Linear Regression Model including interaction for two predictors

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

where

- y is the dependent variable
- x_1, x_2 are the independent (predictor) variables
- $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ is the deterministic part of the model
- $\beta_1 + \beta_3 x_2$ represents the change in y for a 1 unit increase in x_1
- $\beta_2 + \beta_3 x_1$ represents the change in y for a 1 unit increase in x_2
- e is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ.

This means we find for each observation in our sample the product of their measurement for x_1 and times it by the measurement for x_2, this gives us a third predictor variable to be included in the model. We then analyze the data in the same way we did for the first order model.

Continue example:
The model to predict the weight change of snow goslings through their digestion efficiency and the amount of acid-detergent fibre they were fed, including an interaction term is:

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

y = weight change, x_1 = digestion efficiency, and x_2 = acid-detergent fibre
In a first step we find again the estimates for the model parameters (using SPSS)

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	9.561	5.794		1.650	.107	-2.169	21.290
digeff	.024	.090	.106	.270	.788	-.159	.207
adfiber	-.357	.193	-.644	-1.845	.073	-.748	.035
eff-fiber	-.002	.003	-.131	-.702	.487	-.009	.004

a Dependent Variable: wtchng

gives $\hat{\beta}_0 = 9.561$, $\hat{\beta}_1 = 0.024$, $\hat{\beta}_2 = -0.357$, and $\hat{\beta}_3 = -0.002$.
For finding an estimate for σ we need to look at the ANOVA table and find SSE, or MSE

ANOVA(b)

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	548.216	3	182.739	14.561	.000(a)
Residual	476.903	38	12.550		
Total	1025.119	41			

a Predictors: (Constant), adfiber, digeff, eff-fibre
b Dependent Variable: wtchng

We get $MSE = 12.550$, so $s = \sqrt{12.55} = 3.543$.
Now check the utility, for this we will first report the coefficient of determination $R^2 = 0.535$ and the adjusted coefficient of determination $R_a^2 = 0.498$, both indicate that about 50% of the variation in the weight change can be explained through digestion efficiency and the amount of fibre fed.
To assure the model is meaningful, we conduct a model utility test

1. Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : at least one is not zero. Choose α.

2. Assumptions: The assumption of the multiple regression model hold up, we checked that by looking at the scattergram that the model seems to fit reasonably well.

3. Test Statistic from the ANOVA table:

$$F_0 = \frac{(SS_{yy} - SSE)/k}{SSE/(n-k-1)} = 14.561$$

   with $df_n = 3, df_d = 42 - 3 - 1 = 38$.

4. P-value = $P(F > F_0)$=.000

5. Reject H_0 since P-value$< \alpha = 0.05$

6. At significance level of 0.05 the test confirms our conclusion based on the coefficient of determination, we have sufficient evidence that the model is useful in explaining the weight change of snow geese goslings.

In order to decide, if interaction is present, and which variable (beyond the interaction) influences the weight change, we test if the slope parameters (coefficients) are significantly different from 0.

1. Hypotheses: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$, $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$, $H_0 : \beta_3 = 0$ vs. $H_a : \beta_3 \neq 0$

   Choose $\alpha = 0.05$.

2. Assumptions: Regression model is appropriate, checked before.

3. Test statistic (from SPSS, see above):
   For β_1 : $t_0 = 0.270, \quad df = 39$
   For β_2 : $t_0 = -1.845, \quad df = 39$
   For β_3 : $t_0 = -0.702, \quad df = 39$

4. P-value:
   For β_1 : $P - value = 0.788$,
   For β_2 : $P - value = 0.073$
   For β_3 : $P - value = 0.487$

5. Decision: None of the P-values is smaller than 0.05 we can not reject any of the null hypotheses.

6. The tests are non conclusive, we could not find interaction none of the two factors is shown to have a significant influence on the weight change of the goslings.

In our model building process, we should decide, not to include the interaction with the model and go back to the first–order model only including the amount of fibre as a predictor, because digestion efficiency neither showed a significant main effect nor interaction with the amount of fibre on the weight change of gosling when correcting for the influence of the amount of fibre.