Contents

1 Multiple Linear Regression Models

The Multiple Linear Regression Model is introduced as a mean of relating one numerical response variable y to two or more independent (or predictor variables).
This section presents

- different models allowing numerical as well as categorical independent variables
- how to use sample data for obtaining estimates and confidence intervals for the model parameters and how to interpret these
- the assumptions implied by the multiple linear regression model and how to use sample data to check if it reasonable to assume that they are met
- how to apply the model and predict future values and estimate the mean response for given values of the predictor variables

The Multiple Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e, \quad e \sim \mathcal{N}(0, \sigma)$$

where

- y is the response (dependent) variable
- k is the number of predictors
- $x_1, x_2, \ldots, x_k$ are the predictor (independent) variables
- The mean of the response, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, is the deterministic part of the model
- β_i describes the contribution of the predictor variable x_i
- e is the random error, which is assumed to be normally distributed with mean 0 and standard deviation σ.

The following sections introduce choices for the independent variables, when building a model for a response of interest.

1.1 First-Order-Model

The term *first* indicates that the predictor variables are only included in their first power, later higher-order terms will be introduced.

The First-Order Model in numerical Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e, \quad e \sim \mathcal{N}(0, \sigma)$$

where $x_1, x_2, \ldots, x_k$ are independent numerical variables (each variable measures a different concept).
β_0 is the mean of y, when all predictor variables equal 0.

β_i is the slope of the line relating y with x_i when all other independent variables are held fixed. It gives the change in the mean of y, for a 1 unit increase in x_i, leaving all other predictors 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 $\hat{\beta}_1, \ldots \hat{\beta}_k$, where $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$.

Finding the estimates involves solving $k+1$ equations for $k+1$ parameters. This is *relatively easy* with methods from linear algebra, but tedious without.
In this course we will rely on the computer to calculate the estimates. 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 $df_E = n - (k+1)$.

Example 1
Feeding of snow goslings

Botanists were interested in predicting the weight change (in%) of snow goslings as a function of digestion efficiency (in%), percent of acid-detergent fibre in their feed, and diet (plants versus duck chow (duck food)). The acid-detergent fibre is the least digestible portion in plant food.
The goal will be to fit a first order model to the weight change of the goslings:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

y = weight change, x_1 = digestion efficiency, x_2 = acid-detergent fibre.

At this point diet can not be included with the model since it is a categorical variable.

1. **Scatterplots**: In a first step check if a linear model is reasonable choice for describing the relationship between diet efficiency, acid-detergent fibre, and weight change.

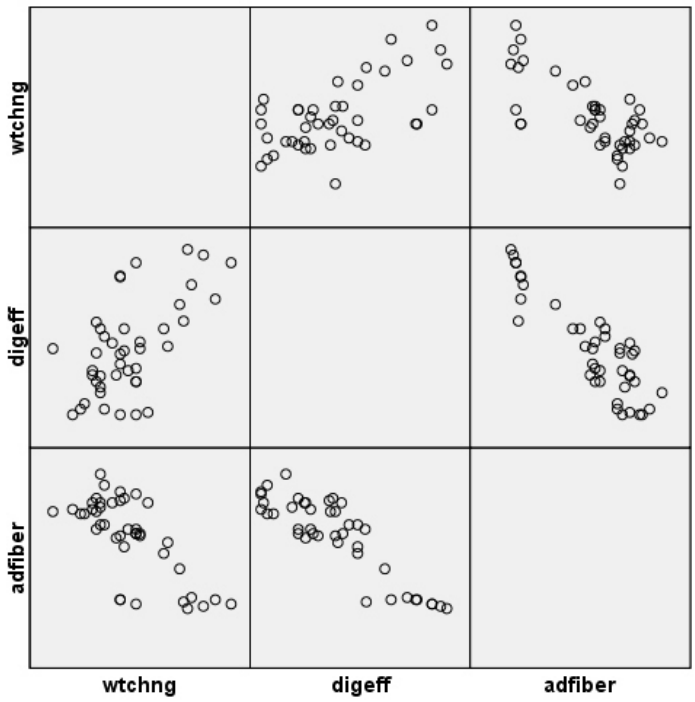
   Obtain scattergrams for

   (a) the response variable with each predictor variable, and

   (b) each pair of predictor variables.

   A linear model can be considered reasonable if the response variable shows a linear relationship with each explanatory variable, or at least does not demonstrate a non-linear relationship, and the scattergrams for the pairs of explanatory variables do *not* show a strong linear relationship.

   To understand this requirement, assume they would relate in a perfect linear relationship. Then in the Multiple Linear Regression Model they would bear the same information on the response variable and one of them would be redundant.



The matrix plot shows that the relationship between weight change and the two predictor variables show a linear trend, which supports the choice of the model. But the scattergram for the predictors, digestion efficiency and fibre, also displays a strong linear relationship, which might imply that one of them is redundant. In the analysis one has to check for signs of this effect.

2. **Estimation**: The estimates of the coefficients and the confidence intervals for those produced by SPSS are:

Coefficients

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B 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:

$$\hat{\beta}_0 = 12.18, \hat{\beta}_1 = -0.027, \text{ and } \hat{\beta}_2 = -0.458$$

resulting in the estimated multiple linear regression function

$$\hat{y} = 12.18 - 0.027x_1 - 0.458x_2$$

The ANOVA table (find the table below)) provides $MSE = 12.387$, which results in $s = \sqrt{12.387} = 3.5195$.

3. **Interpretation:** $\beta_1 = -0.27$ gives the mean change in 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% and the acid-detergent fibre is at 0 %.

Overall Model Utility and Inference about the Parameters

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

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{\text{Explained variability}}{\text{Total variability}}$$

Just as in the simple linear regression model the coefficient of determination measures how much of the sample variability in the response, y, can be explained through the linear model applied.
For the 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 that the model is appropriate for estimation and prediction of the response. R^2 close to zero indicates lack of fit.
One needs to be careful though, because if the sample size equals the number of predictors $R^2 = 1$, 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 predictors when fitting a multiple linear regression model.

In order to adjust for a high number of parameters (predictors) in relation to the sample size, the adjusted-$R^2 = R_a^2$ is used to measure the fit of a multiple linear regression model,

$$R_a^2 = 1 - \frac{n-1}{n-k-1}\left(\frac{SSE}{SS_{yy}}\right)$$

R_a^2 will *not* automatically increase when parameters are added to the model. Smaller models are preferred over larger models.
Usually the adjusted coefficient of determination is reported for multiple linear regression models.
For gosling example $R_a^2 = 0.502$, which is only a little lower than R^2.

Beyond reporting and interpreting R^2 and R_a^2 for assessing model fit, model utility is tested.
If all slope parameters would be 0, then nothing could be learned about the response from the predictor variables in the model. Therefore a test should be conducted to disprove that the slopes are all zero.

Test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ versus H_a : at least one slope is different from 0. The test statistic is based on

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

which is the ratio of the variance in the response explained by the model and the variance remaining unexplained.

A large ratio indicates that most of the variability in y could be explained, and a small value indicates that barely any variability could be explained. To judge if the ratio is small or large the degrees of freedom have to be taken into account.

The F-Test for overall usefulness of the model (Model Utility Test)

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: Random sample, the multiple regression model is an appropriate description of the relationship between predictors and response variables.
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 = $P(F > F_0)$
5. Reject H_0 if P-value$< \alpha$, other wise do not reject H_0.
6. Put into context.

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: Random sample. The assumption of the multiple regression model hold, the scatterplots looked reasonable, but normality of the error and homoscedasticity still needs checking.
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 = $P(F > F_0)$¡.0005
5. Reject H_0 since P-value$< \alpha = 0.05$

6. At significance level of 0.05 the test confirms the conclusion drawn from the coefficient of determination, the data provide sufficient evidence that the model is useful in explaining the weight change of snow goslings.

Once established that the model seems to describe the relationship between the response and the predictor variables, start asking questions about the values of the parameters, and which of them show a significant influence on the response variable.
This question can be answered in two ways, find confidence intervals for the parameters, or conduct statistical tests.

$(1-\alpha)\times 100\%$ **Confidence Interval for a slope parameter** β_i

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

The estimates $\hat{\beta}_i$ and $s_{\hat{\beta}_i}$, should be obtained through SPSS, formulas are beyond the scope of this course.

Continue 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, conclude that β_1 might be 0, and at confidence level of 95% the digestion efficiency has no significant influence on the weight change.
Since the confidence interval for β_2 does not include 0, deduce that at confidence level of 95% $\beta_2 \neq 0$ and in fact $\beta_2 < 0$, that is the less acid-detergent fibre in the food, the more the goslings gain weight.

Another way of answering if individual variables have a significant effect on the mean response is a **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: Test if an increase in acid fibre in the food decreases the weight increase in snow goose goslings when correcting for diet efficiency.

1. Hypotheses: Parameter of interest β_2. the higher x_2 the lower y:

   lower tail $H_0 : \beta_2 \geq 0$ vs. $H_a : \beta_2 < 0$

   Choose $\alpha = 0.05$.

2. Assumptions: checked before.

3. Test statistic: (from first table for this example)

$$t_0 = -3.569, \quad df = 39$$

4. P-value:

   The two-tailed P-value is 0.001, since this is 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, at 5% significance level the data provide sufficient 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.

Residual Analysis
Before applying the model for prediction and estimation check if the assumptions are met.

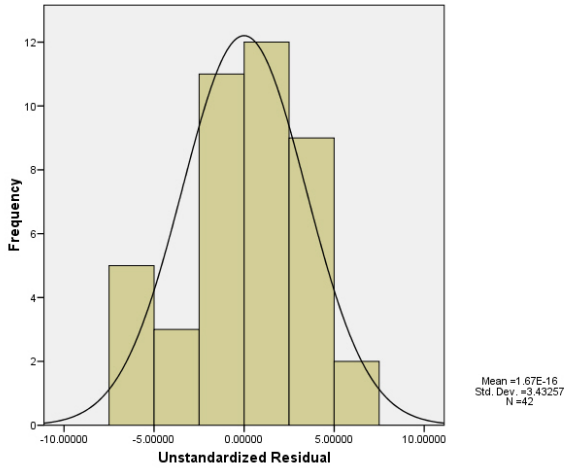
- the error, e, in the regression model is normally distributed with mean 0 and standard deviation σ
- the standard deviation is the same for all values of the predictors.

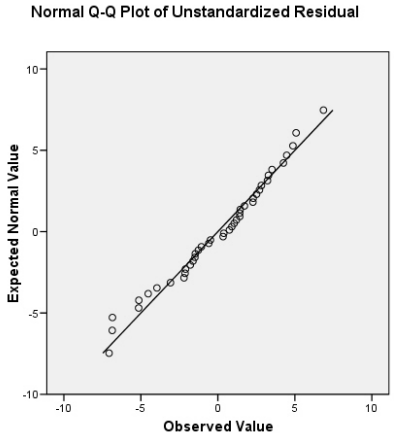
In order to check these assumptions the residuals have to be calculated and

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 example:

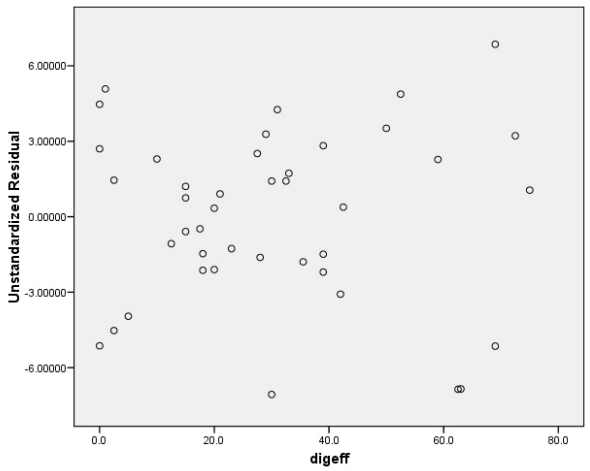
1. The histogram and QQ-plot for the residuals in our example:

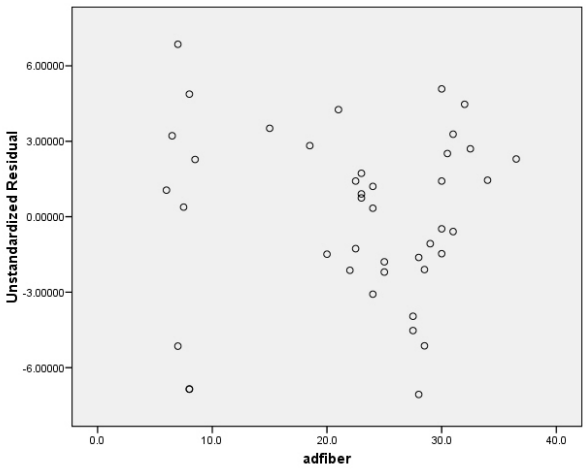




   Both graphs do not indicate a strong deviation from normality. They are no cause for concerns, about the assumptions in the model not being met.

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





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 the conclusion derived so far seem valid.

Now that it has been 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, the model can be used to predict future values of y and estimate $E(y)$ for chosen values of the predictor variables.

Estimation and Prediction Using the Model
Remember the difference between

- estimating the mean of y, $E(y)$, for given values of $x_1, \ldots, x_p$, and
- predicting a future value of y for given values of $x_1, \ldots, x_p$.

For both cases give an interval:

- when estimating $E(y)$, give a confidence interval for the **mean** of y for a chosen level of confidence for chosen values of $x_1, \ldots, x_p$.
- when predicting y, give a prediction interval which shall include a certain proportion of **measurements** of y for the chosen values of $x_1, \ldots, x_p$.

Continue example:
To predict the weight change in a gosling with digestion efficiency of $x_1 = 21\%$ and being fed food including $x_2 = 23\%$ acid-detergent fibre use SPSS (use the Save button within Linear Regression).

The 95% prediction interval (from SPSS) is given to be [-6.16257, 8.34843].
95% of weight changes of goslings with $x_1 = 21$ and x_2=23 fall between -6.2% and 8.3%.

The mean weight change of goslings with 21% digestion efficiency and who are fed 23% acid-detergent fibre is estimated through a 95% CI to fall between -.31 % and 2.5%. This confidence interval indicate that there is no significant mean weight change for snow goslings with 21% digestion efficiency who are fed 23% of acid-detergent fibre.

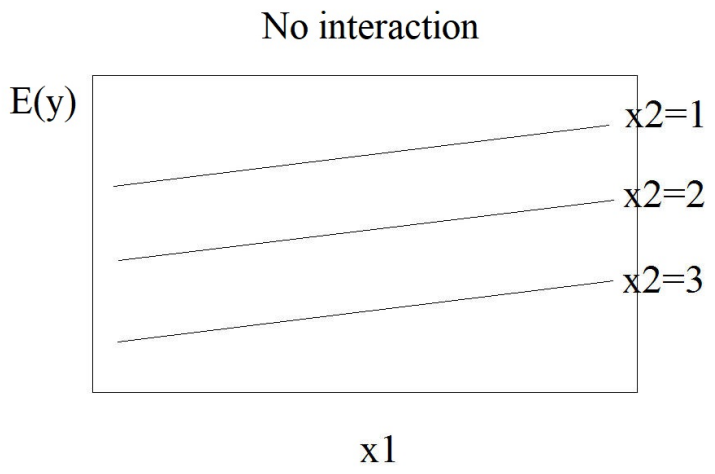
1.2 Interaction Models

Similarly as in ANOVA, the Regression model can be used to investigate if explanatory variables interact in their effect on the response.
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.
Then, when we know the value of $x_2 = x_{2p}$, then

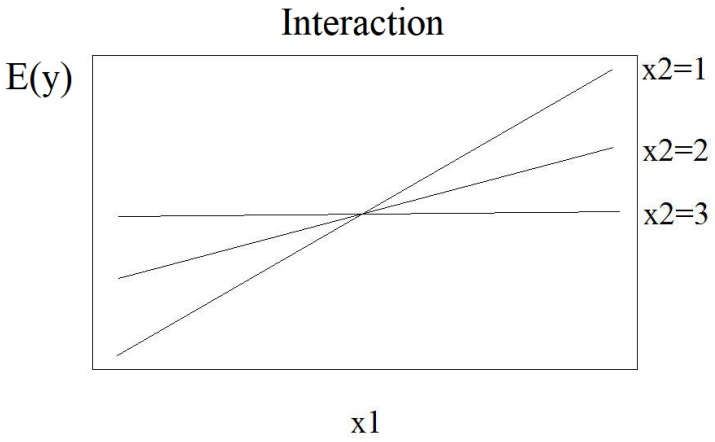
$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_{2p} = \underbrace{\beta_0 + \beta_2 x_{2p}}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} x_1$$

is a line relating x_1 with $E(y)$. The slope of this line is β_1 and intercept is $\beta_0 + \beta_2 x_{2p}$.

In conclusion: $E(y)$ depends on x_1 with the same slope for all values of x_2



The lines for different values of x_2 are parallel!
Including an interaction term with the model will permit to fit lines that are not parallel and to test if a model allowing non-parallel lines provide a significant better fit of the model to the population. The new model (interaction model) allows that the slopes of the lines relating x_1 with y differ for different values of x_2.
This phenomenon indicate that x_1 and x_2 interact in their effect on y. The relationship between x_1 and y depends on the value of x_2.



For example consider adding different additives for watering plants (assume beans).
y = height of plant after 4 weeks
x_1 = concentration of additive
x_2 = additive (sugar, fertilizer, salt)
Depending on what is added to the water the plant will either grow slower, faster, or maybe even die.
The response of the plant to the concentration of the additive depends on the additive.

A Multiple Linear Regression Model including interaction for two predictors

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e, \quad e \sim \mathcal{N}(0, \sigma)$$

where

- y is the dependent variable
- x_1, x_2 are the independent (predictor) variables
- $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ 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.

to introduce interaction in SPSS, a new variable has to be created $x_3 = x_1 x_2$ and included in the model.
For each observation in the sample the product of their measurements for x_1 and x_2 is calculated and this product is used as the third predictor variable in the model.
The analysis is then conducted exactly in the same way as for the first-order mode, but the interpretation of the slopes and test changes.

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e, \quad e \sim \mathcal{N}(0, \sigma)$$

y = weight change, x_1 = digestion efficiency, and x_2 = acid-detergent fibre

The scatterplots have already been assessed.

Find 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 $\hat{\beta}_3 = -0.002$, giving

$$\hat{y} = 9.561 + 0.024 x_1 - 0.357 x_2 - 0.002 x_1 x_2$$

To calculate an estimate for σ find MSE in the ANOVA table:

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

$MSE = 12.550$, so $\hat{\sigma} = s = \sqrt{12.55} = 3.543$.

Next check the utility of the model by interpreting the adjusted R^2 and conducting a model utility test:
Adjusted Coefficient of Determination: $R_a^2 = 0.498$, indicates that about 50% of the variation in the weight change can be explained through digestion efficiency and the amount of fibre fed, and their interaction.

Model Utility Test

1. Hypotheses: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : at least one is not zero. Choose $\alpha = 0.05$.
2. Assumptions: Random sample + residual analysis to check normality and homogeneity.
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 = $P(F > F_0)$¡.0005
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 goslings.

To decide if interaction is present, and which variable (beyond the interaction) influences the weight change, conduct a test for each slope to be significantly different from 0.

1. Hypotheses (test all three simultaneously):
   $H_0 : \beta_1 = 0$ 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: See above.

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.

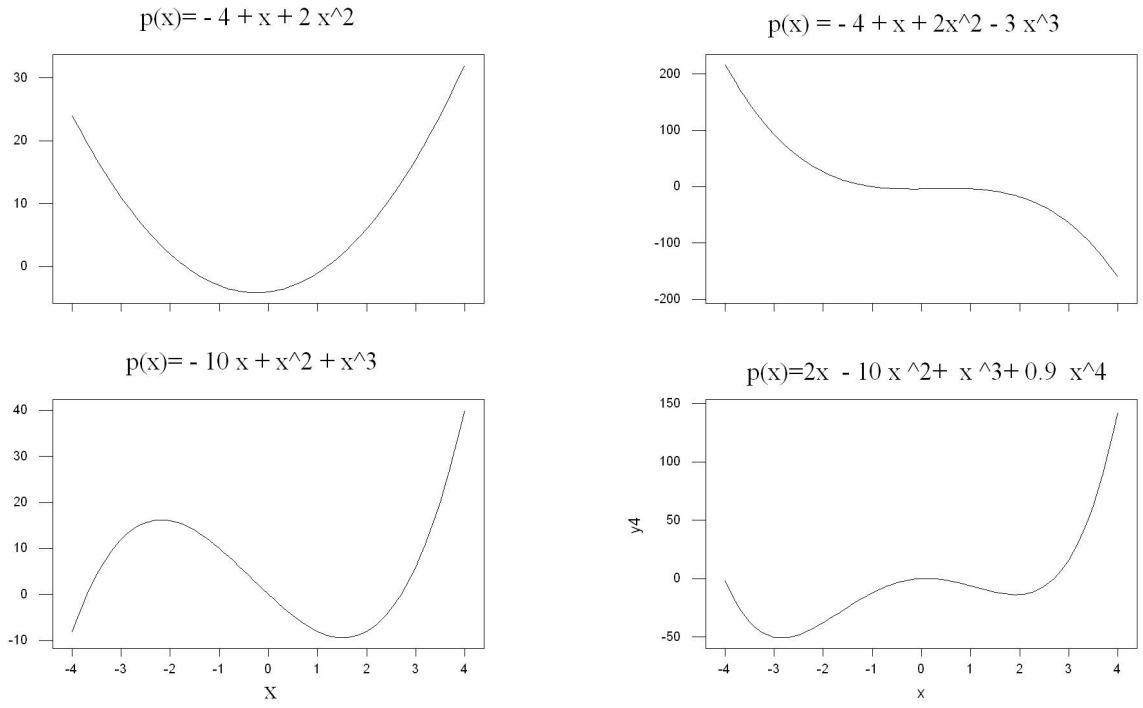
In our model building process, we should decide, not to include the interaction with the model and go back to the first–order model only including the amount of fibre as a predictor, because digestion efficiency neither showed a significant main effect nor interaction with the amount of fibre on the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

1.3 Polynomial (Higher Order) Models

We will now move away from the assumption that the predictor variables and the mean of the response variable share a linear relationship. Instead we will permit that the variables are related through polynomials.
Polynomials of degree k are functions of the form

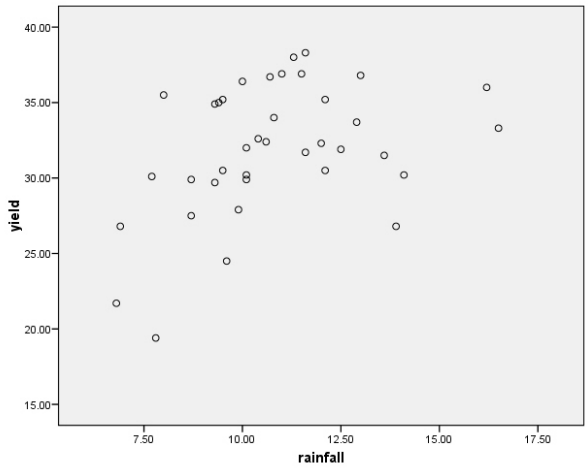
$$p(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k$$

Example 2



When fitting polynomials to data, only part of the function has to fit. For example to model a concave upward relation ship, we can use a polynomial function of degree 2 and only the upward branch of the function will be fitted.

Example 3
In the scattergram below the corn yield (bu/acre) versus rainfall (inches) in six U.S. corn producing states, recorded for each year from 1890 to 1927 is displayed.



A straight line regression model is not adequate. Although increasing rainfall is associated with higher mean yield for rainfalls up to 12 inches, increasing rainfall at higher levels is associated with no change or decrease in the mean yield.
One possible model would be a second order model

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

where y = corn yield and x = rain fall.

For fitting a higher order model we go through the same steps, as before

1. Analyze scattergram, to check if model is adequate.
2. Estimate the parameters of the model (and interpret)
3. Check model utility (R^2, R_a^2, model utility test)
4. Find which factors in the model are relevant.
5. Check model assumptions
6. predict and estimate

Continue Example:

1. We looked at the scattergram before and found that a 2nd order model seems to be appropriate.
2. Obtain a variable rain-sq= x^2. And obtain the estimates with SPSS by including rain and rainsq as predictor in the model.

Coefficients

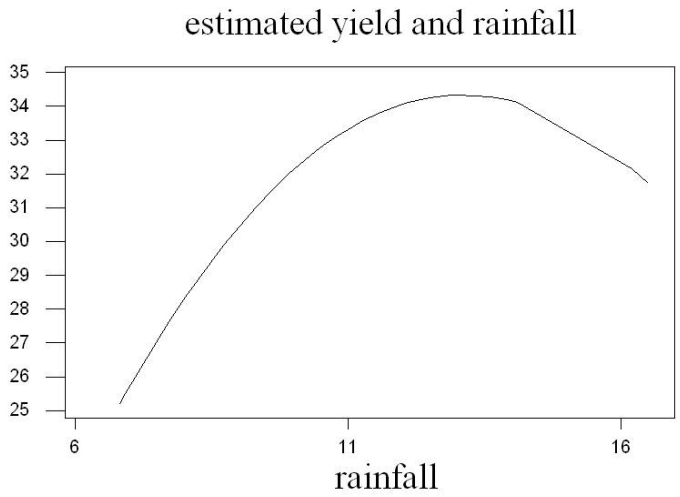
Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
(Constant)	-5.015	11.442		-.438	.664
rainfall	6.004	2.039	3.117	2.945	.006
rainsq	-.229	.089	-2.739	-2.588	.014

Dependent Variable: yield

We find $\hat{\beta}_0 = -5.015, \hat{\beta}_1 = 6.004, \hat{\beta}_2 = -.229$, so $\hat{y} = -5.015 + 6.004x - 0.229x^2$.

The negative estimate for β_2 is consistent with the graph, being concave downward.

The interpretation of $\hat{\beta}_0$ would be the mean yield, for a rainfall of 0 inches, this is not meaningful. Usually the coefficients β_1 and β_2 will have no meaningful interpretation, beyond the comment about the concavity of the quadratic function.



3. The goodness of the fit is measured through $R_a^2 = 0.256$. 25.6% of the variation in the corn yield can be explained by a quadratic model in the rainfall. It indicates that rainfall plays a role in predicting the yield of corn, but other factors should be considered if estimation and prediction shall be tried.

   The model utility test would be
   $H_0 : \beta_1 = \beta_2 = 0$ versus H_a : at least one is not 0. $\alpha = 0.05$

   Assume that the model describes the population, to be checked with a residual analysis.

   Test statistic:

ANOVA(b)

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	209.022	2	104.511	7.382	.002(a)
Residual	495.529	35	14.158		
Total	704.551	37			

a Predictors: (Constant), rainfall, rainsq

b Dependent Variable: yield

$F_0 = 7.382$, $df_n = 2$, $df_d = 35$ and P-value=0.002.

Since the p-value is smaller than the significance level we reject H_0, and decide that at least one of the coefficient is not 0, and the model is a useful predictor for the yield of corn.

4. To check if the model is in fact quadratic or the linear term would have been enough, we test $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$. $\alpha = 0.05$

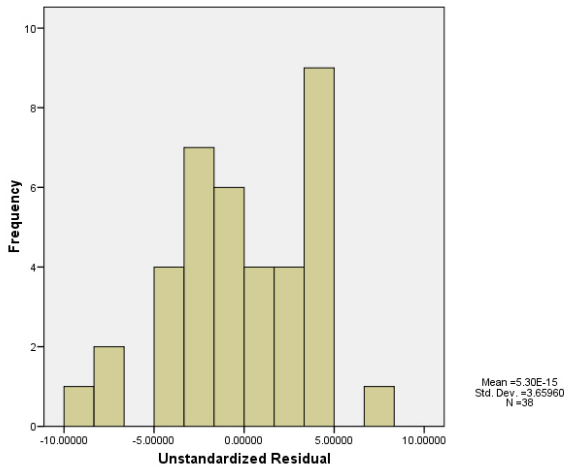
   Assume that the model describes the population, to be checked with a residual analysis.

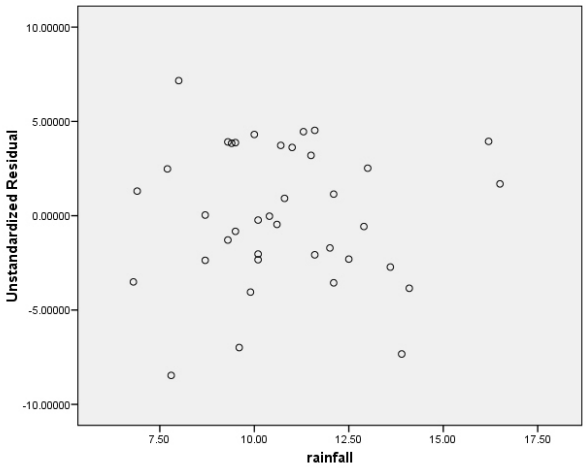
   Test statistic: From the coefficient table we, find $t_0 = -2.588$ and $df = df_E = 35$, with P-value=0.014.

   We reject H_0, and conclude that the quadratic model is appropriate.

   A test for $\beta_1 = 0$ would also be significant, and we conclude that the linear term should be included in the model.

5. For the above conclusion all to be correct the model assumptions have to be met. A residual analysis should be done.





The histogram is not too different from a bellshaped curve to be reason for concern, about the assumption not being met.

And the residual plot shows scatter around 0 of about the same amount, independent from the rainfall, indicating that the assumption that the standard deviation is independent from the rainfall seems to be met.

6. To predict the yield of corn for a rainfall of 13 inches, we find a 95% prediction interval with SPSS: [26.43, 42.13]. In areas with 13 inches of rain, we should expect that 95% of fields yield between 26.43 and 42.13 bu/acre. It is a very large interval judging by the scattergram, reflecting the relatively small R_a^2.

1.4 Including qualitative variables (dummy variables)

So far we studied how with a Multiple Linear Regression Model we can model a numerical response variable in dependency on other numerical predictor variables. Another strength of the model is that we can also include categorical variable with the model, by introducing dummy variables.
To illustrate, consider that we want to model the weight of a person, y, using the height, h, and the gender (male or female).
To do so, we introduce a *dummy variable* using values 0 or 1 for coding the qualitative variable gender.

$$g = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

and propose the model

$$y = \beta_0 + \beta_1 h + \beta_2 g + e$$

assuming that the error e is normally distributed with mean zero and standard deviation σ.
In this model the mean of a person would depend on the gender:

$$\begin{array}{ll} \text{For males } (g = 1): & E(y) = \beta_0 + \beta_1 h + \beta_2 \\ \text{For females } (g = 0): & E(y) = \beta_0 + \beta_1 h \end{array}$$

In the line relating the mean weight of males with their hight the slope is β_1 and the intercept $\beta_0 + \beta_2$.
In the line relating the mean weight of females with their hight the slope is β_1 and the intercept β_0.
(diagram)
In conclusion β_2 is the difference between the mean weight for mean and women, at the same hight.
By testing (based on sample data) if $\beta_2 \neq 0$, we test if the mean weight of men and women is significantly different when correcting for their height. That is, if their mean weight is different, even if their height is the same.

If we want to include a qualitative variable with more than 2 levels, we have to introduce more than 1 dummy variable.

Including one qualitative variable with k levels in a multiple regression model
For including the k levels of the qualitative variable introduce $k - 1$ dummy variables, $d_2, d_3, \ldots, d_k$ (start with 2). Where d_i is the dummy variable for level i.

$$d_i = \begin{cases} 1 & \text{if categorical variable has level i} \\ 0 & \text{otherwise} \end{cases}$$

Then the model, with x being an additional numerical variable, is

$$y = \beta_0 + \beta_1 x + \beta_2 d_2 + \beta_3 d_3 + \ldots + \beta_k d_k + e$$

In this model β_i $(i = 2, \ldots, k)$ is the difference between the mean of y for level i and the mean of y for level 1, when keeping x fixed, because

$$\begin{array}{ll} \text{For level 1:} & E(y) = \beta_0 + \beta_1 x \\ \text{For level } i\text{:} & E(y) = \beta_0 + \beta_1 x + \beta_i \end{array}$$

We could include more numerical variables, interaction terms, or higher order terms with this model.

Example 4
In the snowgeese example an additional variable was included in the data set, being the type of food (chow or plants). This is a categorical variable we will now include in the model. Since the variable has 2 levels, we need to include 2-1=1 dummy variables

$$d_p = \begin{cases} 1 & \text{if plant was fed i} \\ 0 & \text{if chow was fed} \end{cases}$$

We also include $x_1 =$ digestion efficiency and $x_2 = acid - detergentfibre$ with the model to explain $y =$ weight change:

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

Scattergrams have already been produced to rectify the linear model for x_1 and x_2.
For estimating the parameters, the dummy variable has to be created first. In SPSS use Transform>Compute Variable, Target variable: *dp*, and numeric expression: *(diet='Plants')*.
Then as usual the linear regression model is fitted, producing the following coefficient table:

Coefficients

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
(Constant)	12.180	4.460		2.731	.010
digeff	-.026	.054	-.114	-.484	.631
adfiber	-.462	.168	-.834	-2.752	.009
dp	.119	2.869	.010	.041	.967

a Dependent Variable: wtchng

The deterministic part of the model is estimated as $\hat{y} = 12.18 - 0.026x_1 - 0.462x_2 + 0.119d_p$
Estimating that the mean weight change of goslings being fed plants is 0.119% higher than in goslings being fed chow, if the digestion efficiency and the fibre are the same.
The model utility test shows again a useful model ($F_0 = 14.214, df_n 3, df_d = 38$, P-value=0.000) and $Ra^2 = 0.492$.
Testing for influence of the different predictors on the weight change. We find at significance level of 0.05, that only the amount of acid-detergent fibre has an influence on the weight change (if correcting for fibre and diet), but the data do not provide sufficient evidence that digestion efficiency or the diet (plants versus chow) have an impact on the weight change, if correcting for the other factors.
Residual Analysis would establish that the model assumptions are met and our results hold.

In addition we can introduce interaction terms for the dummy variable and the numerical independent variables, which would give us a tool for checking if the slope parameters in the weight change are different for the different diets. Compare with the discussion about interaction.

1.5 Model Building

In the last sections we discussed how to model different type of dependencies within a multiple linear regression model. We went from first-order model, included interaction of 2 numerical variables, included higher order terms to fit curves, and finally saw how to include categorical variables with the model.
The challenge we now face is the process of choosing the "best" model.
In general we prefer parsimonious models, i.e. models that have the least number of variables, but we also want to explain the response variable as good as possible.
In order to achieve this goal, stepwise procedures are proposed.

Stepwise procedure
For a stepwise procedure the user first chooses the response variable and proposes a list of possible predictor variables, including interactions that might be meaningful, higher order terms, and dummy variables, $x_1, \ldots, x_k$.
In a **first** step we now analyze all models with

$$E(y) = \beta_0 + \beta_1 x_i, \quad 1 \leq i \leq k$$

for all k predictor variables, and obtain for each model the test statistic for $H_0 : \beta_1 = 0$ versus $H_0 : \beta \neq 0$.
The variable with the highest absolute value of the test statistic (indicating a significant influence on the response variable) is chosen to be included with the model, call this variable x_1.

In a **second** step, we now search for the model that leads to the best 2 variable model, by adding each of the remaining variables to x_1 in the model:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_i, \quad 2 \leq i \leq k$$

(for all $k-1$ predictor variables not in the model at this point.)
Again we choose to include as the second variable the one that produces the highest absolute value of the test static (indicating a significant influence on the response variable) for testing $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$, when keeping x_1 constant.

In subsequent steps, we check if additional variables can be entered in the model in the same way as in steps one and two.
The procedure stops, when no more significant test statistics can be found.

Example 5
Data from "SPSS, survival manual" 3rd edition, by Julie Pallant
This is a real data file, condensed from a study that was conducted in Australia by students in Educational Psychology. The study was designed to explore the factors that impact on respondents' psychological adjustment and wellbeing. The survey contained a variety of validated scales measuring constructs that the extensive literature on stress and coping suggest influence people's experience of stress. The scales measured self-esteem, optimism, perceptions of control, perceived stress, positive and negative affect, and life satisfaction. A scale was also included that measured people's tendency to present themselves in a favourable or socially desirable manner. The survey was distributed to members of the general public in Melbourne, Australia and surrounding districts. The final sample size was 439, consisting of 42 per cent males and 58 per cent females, with ages ranging from 18 to 82 (mean=37.4).

tpstress	total perceived stress
tpcoiss	total perceived control of internal states
tmast	Total Mastery, High scores indicate higher levels of perceived control over events and circumstances
tmarlow	Total social desirability, score on Marlowe-Crowne scale, which measures the dgree to which people
age	age in years
sex	sex, 1=males, 2=females
smoke	Do you smoke? 1=yes, 2=no

In the example we will explore how well the perceived stress can be predicted by tpcoiss, tmast, tmarlow, age, and sex.