3.5 Hypotheses Tests

After estimating the parameters of the model, we should check if the model is supported by the data, and what can be learnt from the model. We will first focus on what can be learnt from the model (assuming it represents the population well). The first questions to be answered will be if any of the predictors has an effect on the mean response. If evidence shows this to be true, then we should answer the follow-up question, which of the predictors has an effect.

3.5.1 Model Utility Test

To test if any of the predictors has an effect test the null hypothesis that all slopes are 0

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_a : \text{at least on slope is not } 0$$

The test is based on an ANOVA approach, where we analyse the total sum of squares in $\vec{Y}$, and find how much of the variation is due to the variables in the model and how much of the variation remains unexplained.

- The Total Sum of Squares: Let J_n be the $n \times n$ matrix with all entries being 1.

$$SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \vec{Y}'\vec{Y} - \frac{1}{n}\vec{Y}'J_n\vec{Y} = \vec{Y}\,'(I_n - \frac{1}{n}J_n)\vec{Y}$$

- The Residual Sum of Squares (see above):

$$SS_{Res} = \vec{e}'\vec{e} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \vec{Y}'(I_n - H)\vec{Y} = \vec{Y}'\vec{Y} - \hat{\beta}'X'\vec{Y}$$

- The Sum of Squares for Regression:

$$SS_R = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}'X'\vec{Y} - \frac{1}{n}\vec{Y}'J_n\vec{Y} = \vec{Y}\,'(H - \frac{1}{n}J_n)\vec{Y}$$

Comments:

1. $(I_n - \frac{1}{n}J_n)\vec{Y} = (I_n - H)\vec{Y} + (H - \frac{1}{n}J_n)\vec{Y}$
2. $(I_n - H)\vec{Y}$ and $(H - \frac{1}{n}J_n)\vec{Y}$ are orthogonal ($(I_n - H)(H - \frac{1}{n}J_n) = 0$)
3. $SS_T = SS_R + SS_{Res}$ (Pythagoras!)
4. If $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, than $E(SS_R) = k\sigma^2$.
5. $SS_R/\sigma^2 \sim \chi^2$ with $df = k$
6. SS_R and SS_{Res} are independent.

The F-score relating the Sum of Squares for Regression with the Sum of Squares for Residuals is used for the test

$$F = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

Under the assumptions of the MLRM and if $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ this F-score is F-distributed with $df_n = k$ and $df_d = n - k - 1$. (Why?)
The information for this test is usually summarized in an ANOVA table.

Source	SS	df	MS	F
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

ANOVA F-test

1. Hypotheses:

   $$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{versus} \quad H_a : \text{at least on slope is not } 0$$

   Choose α.

2. Assumptions: Random samples, the MLRM is a proper description of the population, including homoscedasticity and normality of the error.

3. Test statistic:

   $$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}},\ df_n = k = p-1,\ df_d = n-k-1 = n-p$$

4. P-value: $P(F > F_0)$

Continue Example. 3.1
Continue the example on the effect of age and weight on blood pressure and the analysis of the model:

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

First find the ANOVA table. We already calculated $SS_{Res} = 5.78$. It is

$$SS_T = \vec{y}'\vec{y} - \frac{1}{n}\vec{y}' J_6 \vec{y} = (120, 141, 124, 126, 117, 129) \begin{pmatrix} 120 \\ 141 \\ 124 \\ 126 \\ 117 \\ 129 \end{pmatrix} - \frac{1}{6}(757)^2 = 354.83$$

Therefore $SS_R = SS_T - SS_{Res} = 349.05$. (Use that $\frac{1}{n}\vec{y}' J_6 \vec{y} = (\sum_{i=1}^n y_1)^2/n$)

ANOVA table

Source	SS	df	MS	F
Regression	349.05	2	174.525	90.58
Residual	5.78	3	1.927	
Total	354.83	5		

ANOVA F-test

1. Hypotheses:
$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{versus} \quad H_a : \text{at least one slope is not 0}$$
Choose $\alpha = 0.05$.

2. Assumptions: Random samples, the MLRM is a proper description of the population, including homoscedacity and normality of the error.

3. Test statistic: $F_0 = 90.58,\ df_n = 2,\ df_d = 3$

4. P-value: $P(F > F_0)$, use F-distribution table from my website, since F_0 is greater than the 0.005 critical value of the F distribution with $df_n = 2$ and $df_d = 3$, the P-value is smaller than 0.005.

5. Therefore the P-value is smaller than α, reject H_0.

6. Context: At significance level of 5% the data provide sufficient evidence that at least one of the slopes is different from zero, i.e. that at least one of age or weight has an effect on the mean blood pressure.

From here arises the question which of the slopes is not zero.

The model utility test does not indicate that the model is a good description for the population, only that if it is a good fit then at least on of the regressors has an effect on the mean response. A residual analysis will help to judge if the model is appropriate for the population under consideration.

Coefficient of determination and adjusted coefficient of determination

- The coefficient of determination is defined as
$$R^2 = 1 - \frac{SS_{Res}}{SS_T} = \frac{SS_R}{SS_T}$$
Since SS_T measures the total variance in the response variable and SS_{Res} measures the amount of variance in the response not accounted for by the model, R^2 gives the proportion of the variation in the response explained by the model.

- The adjusted coefficient of determination, R^2_{adj}, indicates how well the data is represented by the estimated model. In comparison to the coefficient of determination, R^2, it is adjusted for the number of variables in the model. When adding additional variables to an existing model the coefficient of determination will always increase, regardless of the relevance of the variable for the response. The adjusted coefficient of determination includes an adjustment for the

number of variables in the model to correct for this effect, but otherwise is to be interpreted in the same way as R^2.

$$R^2_{adj} = 1 - \frac{MS_{Res}}{MS_T}$$

Continue Example. 3.1 The coefficient of determination for the blood pressure example is

$$R^2_{adj} = 1 - \frac{5.78}{354.83} = 0.98$$

98% of the sample variation in blood pressure can be explained by the changes in weight and age. The adjusted coefficient of determination for the blood pressure example is

$$R^2_{adj} = 1 - \frac{1.927}{354.83/5} = 0.97$$

After penalizing for the extra variable in the model we rather claim that 97% of the sample variation in blood pressure can be explained by the changes in weight and age.

The adjusted coefficient of determination is often used in the model building process.

3.5.2 Test for individual slopes

In order to find out if a variable x_i has a significant linear effect on the mean response when accounting for the other variables in the model, a test for $H_0 : \beta_i = 0,\ 1 \leq i \leq k$ should be conducted.
The test can be based on the t-score standardizing $\hat{\beta}_i$:

Theorem 3.1.
In the MLRM

$$t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 C_{ii}}},$$

where C_{ii} is the i^{th} diagonal entry in $(X'X)^{-1}$, is t-distributed with $df = n - p$.

Conclusion:
Let $i \in \{1, 2, \ldots, k\}$. A $(1-\alpha) \times 100\%$ confidence interval for slope β_i is given by

$$\hat{\beta}_i \pm t^{n-p}_{\alpha/2} \sqrt{\hat{\sigma}^2 C_{ii}},$$

Test for individual slopes given the other regressors in the model:

1. Hypotheses:

Type	Hypotheses	
upper tail	$H_0 : \beta_i \leq \beta_{i0}$ versus	$H_a : \beta_i > \beta_{i0}$
lower tail	$H_0 : \beta_i \geq \beta_{i0}$ versus	$H_a : \beta_i < \beta_{i0}$
2 tail	$H_0 : \beta_i = \beta_{i0}$ versus	$H_a : \beta_i \neq \beta_{i0}$

Choose α.

2. Assumptions: Random samples, the MLRM is a proper description of the population, including homoscedacity and normality of the error.

3. Test statistic:
$$t_0 = \frac{\hat{\beta}_i - \beta_{i0}}{\sqrt{\hat{\sigma}^2 C_{ii}}},\ df = n - p$$

4. P-value:

Type	P-value
upper tail	$P(t > t_0)$
lower tail	$P(t < t_0)$
2 tail	$2P(t > \|t_0\|)$

Continue Example. 3.1
Test if the mean blood pressure increases with increasing weight(x_1) within the proposed model

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

1. Hypotheses: $H_0 : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$, choose $\alpha = 0.05$.

2. Assumptions: Random samples, the MLRM is a proper description of the population, including homoscedasticity and normality of the error.

3. Test statistic:
$$t_0 = \frac{2.38 - 0}{\sqrt{1.927(0.018)}} = 12.78,\ df = 3$$

4. P-value: P-value = $P(t > 12.78)$, using the t-table with $df = 3$, find that the P-value< 0.005.

5. Decision: Reject H_0.

6. Context: At significance level of 5% the data provide sufficient evidence that on average the blood pressure increases with weight when correcting for age.

3.5.3 Test for a subset of slopes

In some applications it is of interest to test if in the presence of some (usually confounding) variables at least one in a set of slopes for different variables is not zero.
Such a question would for example arise, when it is of interest to find the effect of the amount of humour, violence, and drama on the popularity of a movie, after correcting for the effects of age and sex (the confounding variables).

For the development of the test statistic the **extra sum of squares principle** is applied. It is based on the comparison of the residual sum of squares for two models:

1. The full model: model which includes all variables.

2. The reduced model: the model only including the variables to be corrected for.

Essentially it is tested if the inclusion of the extra variables results in a significant decrease in the residual sum of squares.

1. The full model with k predictors, therefore $p = k + 1$ parameters (including the intercept)

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

where $\vec{Y}$ and $\vec{\varepsilon}$ are n dimensional random vectors, the design matrix $X \in \mathbb{R}^{n\times p}$, and the parameter vector $\vec{\beta} \in \mathbb{R}^p$.

To test for the effect of $r < k$ predictors partition the design matrix and the parameter vector such that

$$\vec{Y} = [X_1|X_2] \begin{bmatrix} \vec{\beta}_1 \\ \vec{\beta}_2 \end{bmatrix} + \vec{\varepsilon} = X_1\vec{\beta}_1 + X_2\vec{\beta}_2 + \vec{\varepsilon}$$

with partitioned design matrix where $X_1 \in \mathbb{R}^{n\times(p-r)}$ and $X_2 \in \mathbb{R}^{n\times r}$, and partitioned parameter vector where $\vec{\beta}_1 \in \mathbb{R}^{(p-r)}$ and $\vec{\beta}_2 \in \mathbb{R}^r$.

For the full model with $\hat{\beta} = (X'X)^{-1}X'\vec{Y}$

$$SS_{Res}(\vec{\beta}) = \vec{Y}'\vec{Y} - \hat{\beta}'X'\vec{Y}$$

with $df = n - p$.
We wish to test $H_0 : \vec{\beta}_2 = \vec{0}$.

2. The reduced model, where we assume $\vec{\beta}_2 = \vec{0}$, with $k - r$ predictors and $k - r + 1 = p - r$ parameters:

$$\vec{Y} = X_1\vec{\beta}_1 + \vec{\varepsilon}$$

and with $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'\vec{Y}$

$$SS_{Res}(\vec{\beta}_1) = \vec{Y}'\vec{Y} - \hat{\beta}_1'X_1'\vec{Y}$$

with $df = n - (p - r) = n - p + r$.

The difference in the Sum of Squares for Residuals when adding variables $x_{k-r+1}, \ldots, x_k$ to the model with variables $x_1, \ldots, x_{k-r}$ is

$$SS_{Res}(\vec{\beta}_2|\vec{\beta}_1) = SS_{Res}(\vec{\beta}_1) - SS_{Res}(\vec{\beta})$$

with $df = n - p + r - (n - p) = r$. These are called the extra sum of squares due to $\vec{\beta}_2$ (extra sum of squares for regression, since the decrease in the residual sum of squares implies an increase by the same amount in the regression sum of squares).
One can prove that $SS_{Res}(\vec{\beta}_2|\vec{\beta}_1)$ is independent of MS_{Res} therefore

$$F = \frac{SS_{Res}(\vec{\beta}_2|\vec{\beta}_1)/r}{MS_{Res}}$$

is F distributed with $df_n = r$ and $df_d = n - p$.

ANOVA Extra Sum of Squares Test:

1. Hypotheses:

$$H_0 : \beta_{k-r+1} = \cdots = \beta_k = 0 (\vec{\beta}_2 = \vec{0}) \quad \text{versus} \quad H_a : \text{at least one slope is not } 0$$

Choose α.

2. Assumptions: Random samples, the MLRM is a proper description of the population, including homoscedasticity and normality of the error.

3. Test statistic:

$$F_0 = \frac{SS_{Res}(\vec{\beta}_2 | \vec{\beta}_1)/r}{MS_{Res}}, \ df_n = r, \ df_d = n - k - 1$$

4. P-value: $P(F > F_0)$

Continue Example. 3.1
The model is

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Use the extra sum of squares test to see if age has a significant effect on the mean blood pressure when correcting for weight.

- The full model:

$$bp = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- The reduced model:

$$bp = \beta_0 + \beta_1(weight) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

1. Hypotheses:

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_a : \beta_2 \neq 0$$

Choose α.

2. Assumptions: Random samples, the MLRM is a proper description of the population, including homoscedasticity and normality of the error.

3. Test statistic: $r = 1$. From previous work $SS_{Res}(\vec{\beta}) = 5.78$ and $MS_{Res} = 1.927$.
For finding $SS_{Res}(\beta_1)$

$$X_1 = \begin{pmatrix} 1 & 69 \\ 1 & 83 \\ 1 & 77 \\ 1 & 75 \\ 1 & 71 \\ 1 & 73 \end{pmatrix} \quad \text{with} \quad X_1'X_1 = \begin{pmatrix} 6 & 448 \\ 448 & 33574 \end{pmatrix} \quad \text{and} \quad (X_1'X_1)^{-1} = \begin{pmatrix} 45.370 & -0.605 \\ -0.605 & 0.008 \end{pmatrix}$$

and

$$X_1'\vec{y} = \begin{pmatrix} 757 \\ 56705 \end{pmatrix} \quad \text{gives} \quad SS_{Res}(\beta_1) = \vec{y}'\vec{y} - (X_1\vec{y})'(X_1'X_1)^{-1}(X_1\vec{y}) = 85.276$$

$$F_0 = \frac{(85.28 - 5.78)/1}{1.927} = 41.4, \ df_n = r = 1, \ df_d = n - p = 3$$

4. P-value: $0.005 < P(F > F_0) < 0.01$ according to the F-distribution table.

5. Decision: Reject H_0.

6. Context: At significance level of 5% the data provide sufficient evidence that age has a significant effect on the blood pressure when correcting for weight.

R

```
reduced<-lm(BP~WEIGHT)
full<-lm(BP~WEIGHT+AGE)
anova(reduced, full)
```

3.6 Orthogonal Columns in the Design Matrix

Definition 3.1.
Matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{n\times k}$ are orthogonal, iff $A'B = 0_{m\times k}$.

Theorem 3.2.
If the design matric can be partitioned, so that

$$X = [X_1|X_2] \text{ with } X_1'X_2 = 0$$

then the least squares estimator for $\vec{\beta} = (\vec{\beta_1}', \vec{\beta_2}')'$ can be partitioned into $\hat{\beta} = (\hat{\beta}_1', \hat{\beta}_2')'$, so that

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1\vec{Y}, \text{ and } \hat{\beta}_2 = (X_2'X_2)^{-1}X_2\vec{Y}$$

Proof:
It is $\hat{\beta} = (X'X)^{-1}X'\vec{Y}$. First find $(X'X)^{-1}$:

$$X'X = \left[\frac{X_1'}{X_2}\right][X_1|X_2] = \left[\begin{array}{c|c} X_1'X_1 & X_1'X_2 \\ \hline X_2'X_1 & X_2'X_2 \end{array}\right] = \left[\begin{array}{c|c} X_1'X_1 & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & X_2'X_2 \end{array}\right]$$

Then

$$A = \left[\begin{array}{c|c} (X_1'X_1)^{-1} & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & (X_2'X_2)^{-1} \end{array}\right]$$

is the inverse of $X'X$, because

$$\begin{aligned} X'XA &= \left[\begin{array}{c|c} X_1'X_1 & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & X_2'X_2 \end{array}\right]\left[\begin{array}{c|c} (X_1'X_1)^{-1} & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & (X_2'X_2)^{-1} \end{array}\right] \\ &= \left[\begin{array}{c|c} (X_1'X_1)(X_1'X_1)^{-1} + 0_{(p-r)\times r}0_{r\times(p-r)} & (X_1'X_1)0_{(p-r)\times r} + 0_{r\times(p-r)}(X_2'X_2)^{-1} \\ \hline 0_{r\times(p-r)}(X_1'X_1)^{-1} + (X_2'X_2)0_{r\times(p-r)} & 0_{r\times(p-r)}0_{(p-r)\times r} + (X_2'X_2)(X_2'X_2)^{-1} \end{array}\right] \\ &= \left[\begin{array}{c|c} I_{p-r} & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & I_r \end{array}\right] = I_p \end{aligned}$$

With this result now find $\hat{\beta}$

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'\vec{Y} \\
&= \left[\begin{array}{c|c} (X_1'X_1)^{-1} & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & (X_2'X_2)^{-1} \end{array}\right] \left[\begin{array}{c} X_1' \\ \hline X_2 \end{array}\right] \vec{Y} \\
&= \left[\begin{array}{c|c} (X_1'X_1)^{-1} & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & (X_2'X_2)^{-1} \end{array}\right] \left[\begin{array}{c} X_1'\vec{Y} \\ \hline X_2\vec{Y} \end{array}\right] \\
&= \left[\begin{array}{c} (X_1'X_1)^{-1}X_1\vec{Y} \\ \hline (X_2'X_2)^{-1}X_2'\vec{Y} \end{array}\right] \\
&= \left[\begin{array}{c} \hat{\beta}_1 \\ \hline \hat{\beta}_2 \end{array}\right]
\end{aligned}$$

q.e.d.

Conclusion: Because of the theorem, we find in the case of orthogonal columns in the design matrix, that the parameter vector can be partitioned, so that the parts are independent from each other, and can be found independently. The extra sum of squares test for $H_0 : \vec{\beta}_2 = 0$ is in the case of orthogonal columns the same as the model utility test in the partitioned model $\vec{Y} = X_2\vec{\beta}_2 + \vec{\varepsilon}$.

For the full model, it is now easy to prove that

$$SS_R(\vec{\beta}) = SS_R(\vec{\beta}_1) + SS_R(\vec{\beta}_2)$$

I.e. the sum of squares for regression for the complete parameter vector equals the sum of squares for regression for the partitioned parameter vector, and therefore according to the extra sum of squares principle the two are independent, and tests for one partitioned vector is independent from the other.