Statistik III

Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

Multiple Regression

Introduction

- In multiple regression a (metric) response variable Y (DV) is predicted by a function of p (arbitrary scaled) predictor variables (IVs) $X_1, \ldots, X_p$
- Approach
  - The ideal situation would be a perfect prediction of Y by a function f

$$Y = f(X_1, X_2, X_3, \ldots, X_p)$$

  - As we will not be able to measure all possible influences, in general we will allow for a prediction like

$$Y = f(X_1, X_2, X_3, \ldots, X_p) + \varepsilon$$

where ε represents unknown/forgotten predictors, random mechanisms, disturbances, ...

Multiple Regression

Introduction

- Approach
  - Without any further knowledge, we initially choose a simple function f, such as a linear function

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p + \varepsilon$$

  - If we have further information on the relationship (e. g., quadratic, exponential, ...) then it should be implemented instead

Example: Cheddar Cheese

- 30 samples of cheddar cheese were analyzed for their chemical composition and were subjected to taste tests
- Measured variables were the taste quality (Y) and the concentration of acetic acid (X_1), hydrogen sulfide (X_2), and lactic acid (X_3)

Variant	Y	X_1	X_2	X_3
1	12.3	4.543	3.135	0.86
2	20.9	5.159	5.043	1.53
3	39.0	5.366	5.438	1.57
4	47.9	5.759	7.496	1.81
⋮				
30	5.5	6.176	4.787	1.25

- Questions
  - Can we predict the taste quality from the ingredients (chemical composition)?
  - Which of the variables is most important for the prediction?

Model

Notation

- Let $\mathbf{y}$ denote the random vector resulting from n observations of the random variable Y

$$\mathbf{y} = \begin{pmatrix} Y_1 & Y_2 & \dots & Y_n \end{pmatrix}'$$

- The random matrix $\mathbf{X}$ denotes the result of n observations of the random vector $\mathbf{x} = (X_1, \dots, X_p)'$, where a vector of 1s is inserted as the first column

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}$$

- Inserting the column vector of 1s in $\mathbf{X}$ allows for a closed representation of the regression equation

Model

Notation

- Let $\boldsymbol{\beta}$ denote the $(p+1)$-vector of unknown parameters (regression coefficients)

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \dots & \beta_p \end{pmatrix}'$$

  where β_0 is interpreted as intercept and $\beta_1, \dots, \beta_p$ as regression weights w.r.t. the random variables $X_1, \dots, X_p$

- $\boldsymbol{\varepsilon}$ is the n-vector capturing the unobserved random error related to the n observations

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \dots & \varepsilon_n \end{pmatrix}'$$

Model

Model equation

- Prediction of Y_i via the predictor variables $X_{i1}, \ldots, X_{ip}$ $(i = 1, \ldots, n)$ is based on the equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

- So we obtain the system of equations

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_p X_{1p} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_p X_{2p} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_p X_{np} + \varepsilon_n \end{aligned}$$

Model

Model equation

- In matrix and vector notation, we write

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or, respectively,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Model: Assumptions

- For the n-vector ε we assume homoscedasticity (homogeneity of variances)
  - All of the components ε_i have the same variance σ^2
  - Components ε_i and ε_j with $i \neq j$ are uncorrelated

$$Cov(\varepsilon) = \sigma^2 \mathbf{I}_n = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

- The expected value of the error is zero

$$E(\varepsilon) = \mathbf{0}$$

- ε and $\mathbf{X}$ are stochastically independent (follows from assuming a joint normal distribution for Y and $X_1, \dots, X_p$)
- For tests and confidence intervals we additionally assume that ε follows a multivariate normal distribution

$$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad \text{or} \quad \varepsilon \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Parameter Estimation

Data

- Considering n realizations of the given random variables in a sample (of size n), we obtain the equation

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{e}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}}$

Parameter Estimation

Estimation procedure

- Estimation of the coefficients $\boldsymbol{\beta}$ in

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

  is based on an ordinary least squares (OLS) criterion, as in simple linear regression
- This requires to minimize the sum of squares

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})]^2 \end{aligned}$$

Parameter Estimation

Estimation procedure

- Minimization is achieved by differentiating

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

with respect to $\boldsymbol{\beta}$, setting the derivative to zero and solving the resulting system of equations

- It can be shown that any estimate $\hat{\boldsymbol{\beta}}$ minimizing this expression satisfies the *normal equation*

$$\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

Parameter Estimation

Estimation procedure

- Under the assumption that the design matrix $\mathbf{X}$ has full rank (for $n \gg p$ we have $rk(\mathbf{X}) = p + 1$) also $\mathbf{X}'\mathbf{X}$ has full rank
- It follows that $\mathbf{X}'\mathbf{X}$ is invertible and we can find unique solutions to the normal equations
- The least-squares estimator and its covariance matrix then are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- An estimate of the error variance σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\mathbf{e}}'\hat{\mathbf{e}} = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

Example: Parameter Estimation

- Let's consider the regression of a random variable Y on two predictors X_1 and X_2

Data

i	y_i	x_{i1}	x_{i2}
1	3	2	1
2	2	3	5
3	4	5	3
4	5	7	6
5	8	8	7

System of equations

$$3 = \beta_0 + 2\,\beta_1 + 1\,\beta_2 + e_1$$
$$2 = \beta_0 + 3\,\beta_1 + 5\,\beta_2 + e_2$$
$$4 = \beta_0 + 5\,\beta_1 + 3\,\beta_2 + e_3$$
$$5 = \beta_0 + 7\,\beta_1 + 6\,\beta_2 + e_4$$
$$8 = \beta_0 + 8\,\beta_1 + 7\,\beta_2 + e_5$$

Matrix notation

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{e}$$
$$\begin{pmatrix}3\\2\\4\\5\\8\end{pmatrix} = \begin{pmatrix}1&2&1\\1&3&5\\1&5&3\\1&7&6\\1&8&7\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\\\beta_2\end{pmatrix} + \begin{pmatrix}e_1\\e_2\\e_3\\e_4\\e_5\end{pmatrix}$$

Example: Parameter Estimation

- In order to determine $\hat{\boldsymbol{\beta}}$, we need

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \\ 1 & 5 & 3 & 6 & 7 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \\ 1 & 8 & 7 \end{pmatrix} = \begin{pmatrix} 5 & 25 & 22 \\ 25 & 151 & 130 \\ 22 & 130 & 120 \end{pmatrix}$$

and

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 7 & 8 \\ 1 & 5 & 3 & 6 & 7 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{pmatrix} = \begin{pmatrix} 22 \\ 131 \\ 111 \end{pmatrix}$$

Example: Parameter Estimation

- Computing the inverse of $\mathbf{X}'\mathbf{X}$ we obtain

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{1016}\begin{pmatrix} 1220 & -140 & -72 \\ -140 & 116 & -100 \\ -72 & -100 & 130 \end{pmatrix} \quad \text{with} \quad \det(\mathbf{X}'\mathbf{X}) = 1016$$

- Plugging this into

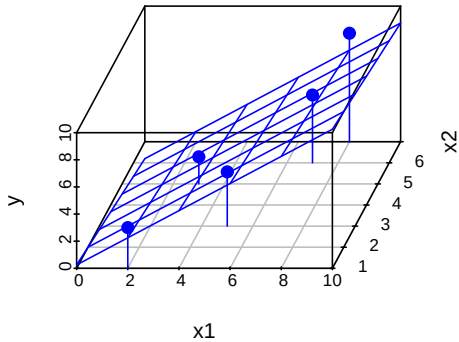
$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

provides the estimate

$$\hat{\boldsymbol{\beta}} = \frac{1}{1016}\begin{pmatrix} 1220 & -140 & -72 \\ -140 & 116 & -100 \\ -72 & -100 & 130 \end{pmatrix}\begin{pmatrix} 22 \\ 131 \\ 111 \end{pmatrix} = \begin{pmatrix} 0.50 \\ 1.00 \\ -0.25 \end{pmatrix}$$

Example: Parameter Estimation

Visualization of the regression equation



Interpretation

- If X_1 increases by 1 unit and X_2 remains constant, then Y increases by 1 unit ($\hat{\beta}_1 = 1$)
- If X_2 increases by 1 unit and X_1 remains constant, then Y drops by 0.25 units ($\hat{\beta}_2 = -0.25$)

Prediction

- Prediction of the response variable (Y) is based on the estimated regression coefficients via

$$\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$$

- The vector $\mathbf{x}_i$ may be a data point that was not observed, and thus not used in estimation
- In our example, Y is predicted from the observations $i = 1, \ldots, 5$ through the regression equation

$$\begin{aligned} \hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}} &= (1 \; x_{i1} \; x_{i2}) \begin{pmatrix} 0.50 \\ 1.00 \\ -0.25 \end{pmatrix} \\ &= 0.50 + x_{i1} - 0.25\, x_{i2} \end{aligned}$$

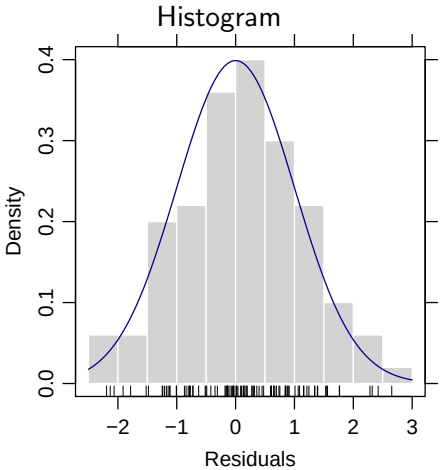
Checking Model Assumptions

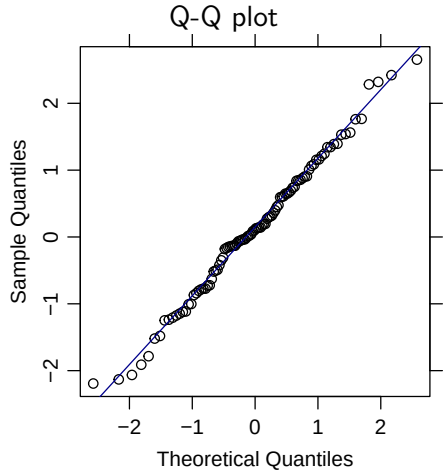
Graphical methods

- The model relies on the assumptions of linearity, independent normally distributed errors, and homogeneous variances (homoscedasticity)
- Plausibility of these assumptions can be tested by checking the residuals in appropriate graphs
  - The assumption of normal distribution may be checked with histograms or Q-Q (quantile-quantile) plots of the residuals
  - The assumptions of linearity and homogeneity of variances may be checked with scatter plots of the residuals vs. the predicted values
- Both procedures may not always provide a clear indication (especially for small sample sizes)

Checking Model Assumptions

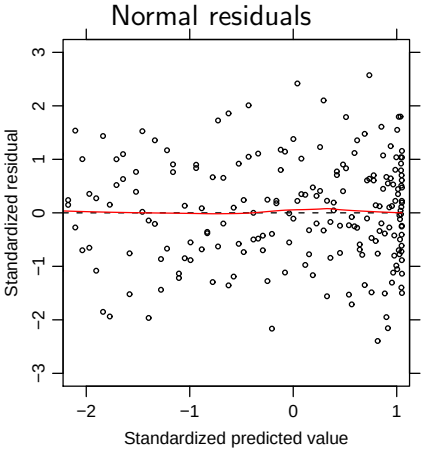
Normally distributed residuals ($n = 100$)

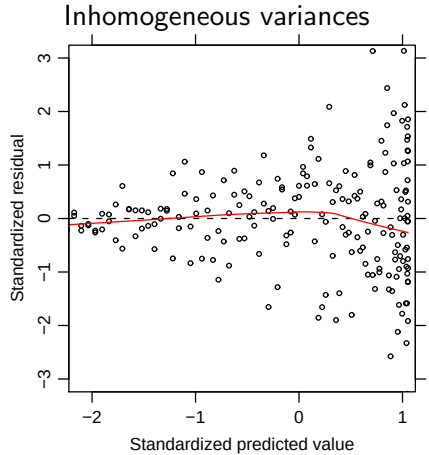




Checking Model Assumptions

Prototypical patterns

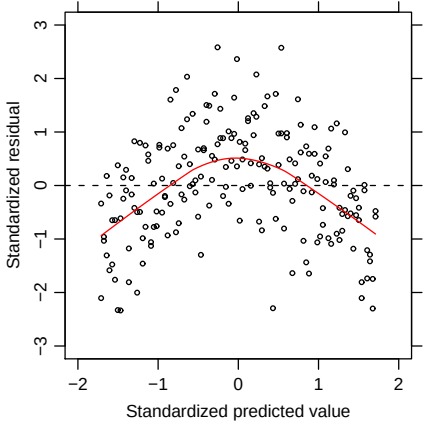




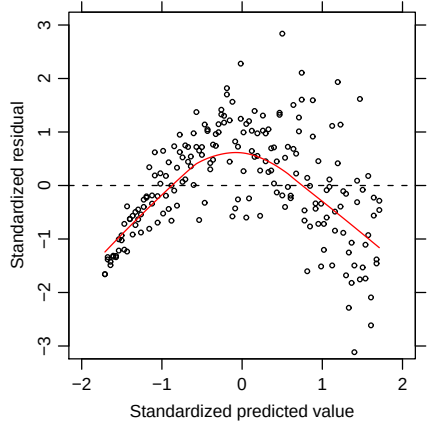
Checking Model Assumptions

Prototypical patterns

Non-linear dependencies



Inhomogeneity + non-linearity



Checking Model Assumptions

Leverage

- *Leverage* measures the distance of the predictor values of a single observation to the remaining observations
- In simple regression (1 IV) it is the proportion of the squared deviation of the observation x_i from the mean $\bar{x}$ to the total sum of all squared deviations

$$l_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- We have $\sum_{i=1}^{n} l_i = 1$, and l_i determines the relative weight with which observation i enters into the computation of the regression line

Checking Model Assumptions

Leverage

- In multiple regression (> 1 IV), the *hat matrix*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

  is calculated
- The name derives from

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

- The leverage is given by the diagonal elements h_{ii} $(i = 1, \ldots, n)$ of the matrix $\mathbf{H}$
- Since the h_{ii} sum to $p + 1$ (p IVs $+ 1$ intercept), and there are n of them, we speak of a high leverage whenever $h_{ii} > 2(p + 1)/n$

Example: Checking Model Assumptions

- Plugging in the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ computed above, we obtain the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \frac{1}{508}\begin{pmatrix} 355 & 67 & 201 & -20 & -95 \\ 67 & 477 & -93 & 32 & 25 \\ 201 & -93 & 229 & 96 & 75 \\ -20 & 32 & 96 & 180 & 220 \\ -95 & 25 & 75 & 220 & 283 \end{pmatrix}$$

- The diagonal elements

$$h_{11} = 0.699, \quad h_{22} = 0.939, \quad h_{33} = 0.451,$$
$$h_{44} = 0.354, \quad h_{55} = 0.557$$

characterize the leverages (all of them $< 6/5 = 1.2$)

Checking Model Assumptions

Influential observations

- We call an observation *influential* if eliminating it from the data leads to a substantial change of the estimated regression equation
- Cook's distance provides a measure of the influence of an observation, and is defined by

$$\frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{(i)j})^2}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(p+1)\hat{\sigma}^2}$$

where $\hat{y}_{(i)j}$ is the predicted value for the jth observation and $\hat{\boldsymbol{\beta}}_{(i)}$ is the estimated parameter vector with observation i deleted from the data

Checking Model Assumptions

Influential observations

- We can compute Cook's distance of observation i via the hat matrix

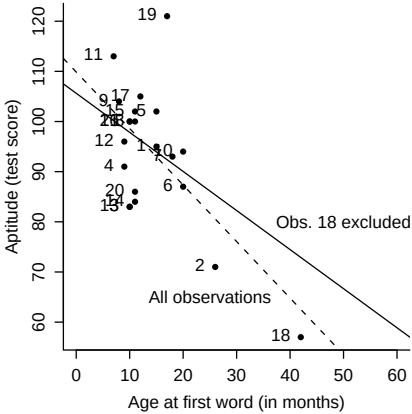
$$\frac{(y_i - \hat{y}_i)^2}{(p+1)\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

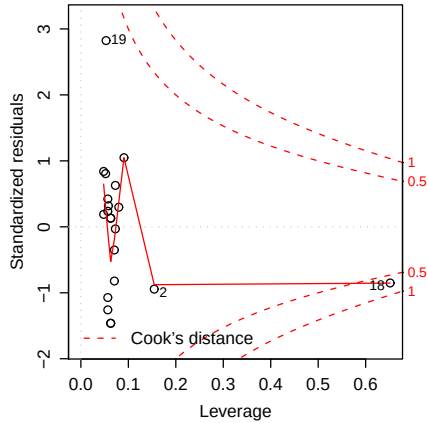
- There are different rules of thumb for a Cook's distance that qualifies an observation as influential
  - A general suggestion is to examine observations with Cook's distance > 0.5, and perhaps eliminate them
  - The formal equivalence of Cook's distance to the Wald test rather suggests the quantile $F_{1-\alpha}(p+1, n-p-1)$ as threshold (known to be a conservative criterion)

Checking Model Assumptions

Influential observations

- Relationship of language development and aptitude (Gesell Adaptive Score) according to Mickey, Dunn & Clark (1967)





Overall Significance Test

Objectives

- The overall significance test (overall F-test) examines whether the predictor variables $X_1, \ldots, X_p$ contribute at all to explaining the response variable Y
- It tests the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{or} \quad H_0 : \; y_i = \beta_0 + \varepsilon_i$$

  versus the alternative hypothesis

$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, p\}$$

- As in simple regression, under the assumption of the null hypothesis with $\beta_0 = \bar{y}$, we consider the model $\hat{\mathbf{y}} = \bar{\mathbf{y}}$

Overall Significance Test

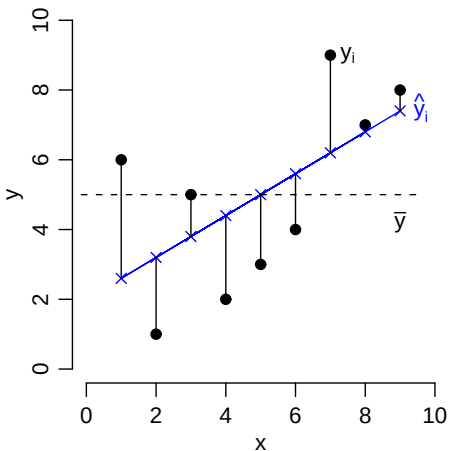
Variance decomposition

- We start from the additive decomposition of the total variability under the null hypothesis into the variability explained by the regression and the residual variability

$$
\begin{aligned}
SS_{tot} &= SS_{reg} + SS_{res} \\
(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) &= (\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}}) + (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2
\end{aligned}
$$

Overall Significance Test

Variance decomposition



$$SS_{tot} = SS_{reg} + SS_{res}$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Coefficient of Determination

- The coefficient of determination R^2 characterizes how well the model is able to reproduce the observations
- R^2 equals the ratio of the regression sum of squares (SS_{reg}) to the total sum of squares (SS_{tot})

$$R^2 = \frac{(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})} = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{\mathbf{e}'\mathbf{e}}{SS_{tot}}$$

- R^2 is to be interpreted as the proportion of variance explained by the regression
- The square root R of the coefficient of determination denotes the multiple correlation between observations $\mathbf{y}$ and predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

Example: Coefficient of Determination

- The predictions of the regression are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & 5 \\ 1 & 5 & 3 \\ 1 & 7 & 6 \\ 1 & 8 & 7 \end{pmatrix} \begin{pmatrix} 0.50 \\ 1.00 \\ -0.25 \end{pmatrix} = \begin{pmatrix} 2.25 \\ 2.25 \\ 4.75 \\ 6.00 \\ 6.75 \end{pmatrix}$$

and the mean is

$$\bar{y} = 4.4$$

with $\mathbf{y} = (3, 2, 4, 5, 8)'$

Example: Coefficient of Determination

- For the regression sum of squares we obtain

$$
\begin{aligned}
SS_{reg} &= (\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}}) \\
&= (-2.15, -2.15, 0.35, 1.60, 2.35) \begin{pmatrix} -2.15 \\ -2.15 \\ 0.35 \\ 1.60 \\ 2.35 \end{pmatrix} \\
&= 17.45
\end{aligned}
$$

Example: Coefficient of Determination

- For the total sum of squares we obtain

$$\begin{aligned} SS_{tot} &= (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) \\ &= (-1.4, -2.4, -0.4, 0.6, 3.6) \begin{pmatrix} -1.4 \\ -2.4 \\ -0.4 \\ 0.6 \\ 3.6 \end{pmatrix} \\ &= 21.2 \end{aligned}$$

- Then the residual sum of squares is given by

$$SS_{res} = SS_{tot} - SS_{reg} = 21.2 - 17.45 = 3.75$$

Example: Coefficient of Determination

- The coefficient of determination is computed as

$$R^2 = \frac{(\hat{\mathbf{y}} - \bar{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{\mathbf{y}})}{(\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}})} = \frac{17.45}{21.2} = 0.823$$

- The correlation between predicted and observed values of the dependent variable then is

$$R = 0.907$$

Overall Significance Test

Test statistic

- An F-distributed test statistic can be derived from the sums of squares as in ANOVA
- The degrees of freedom equal the number p of predictors for the regression sum of squares SS_{reg}, and $n - p - 1$ (with n the sample size) for the residual sum of squares SS_{res}
- Dividing the sums of squares by their respective degrees of freedom results in the mean sums of squares (MS)

$$\frac{SS_{reg}}{p} \quad \text{and} \quad \frac{SS_{res}}{n-p-1}$$

Overall Significance Test

Test statistic

- The ratio of the mean sums of squares

$$F = \frac{\frac{SS_{reg}}{p}}{\frac{SS_{res}}{n-p-1}}$$

is an $F(p, n-p-1)$-distributed random variable (with p and $n-p-1$ degrees of freedom)

Test decision

- The null hypothesis is rejected for given α if

$$F > F_{1-\alpha}(p, n-p-1)$$

Overall Significance Test

Test statistic

- The test statistic can also be computed from the coefficient of determination R^2

$$F = \frac{\frac{R^2}{p}}{\frac{1-R^2}{n-p-1}} = \frac{\frac{SS_{reg}}{p}}{\frac{SS_{res}}{n-p-1}}$$

- The overall F-test is not to be conceived as a goodness-of-fit test, indicating whether the relationship can be described adequately by the linear model
- It merely tests if a linear combination of the predictor values substantially contributes to the prediction of the dependent variable

Example: Overall Significance Test

- We have the following results

Source	*SS*	*df*	*MS*	*F*
Regression	17.45	2	8.725	4.653
Residuals	3.75	2	1.875	
$\sum$	21.20	4		

- We do not reject the null hypothesis (for $\alpha = 0.05$), because

$$F < F_{0.95}(2, 2) = 19$$

- This means that the data do not allow us to conclude any of the predictors contribute to the prediction of Y

Testing Individual Regression Coefficients

Objectives and hypothesis

- This test checks whether the jth independent variable X_j linearly influences the response variable while keeping the other predictors constant (conditional effect)
- It tests the null hypothesis

$$H_0 : \beta_j = 0$$

versus the alternative hypothesis

$$H_1 : \beta_j \neq 0$$

Testing Individual Regression Coefficients

Covariance matrix of $\hat{\boldsymbol{\beta}}$

- It can be shown that the covariance matrix of the estimated parameters is given by

$$Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- This requires an estimate of the unknown error variance σ^2, which is obtained by the mean residual sum of squares

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-p-1} = \frac{1}{n-p-1}(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})$$

- Computation of the test statistic draws upon the jth diagonal element of the $((p+1)\times(p+1))$-matrix $(\mathbf{X}'\mathbf{X})^{-1}$, denoted by a_{jj} (in this case the index j runs over $0,\ldots,p$)

Testing Individual Regression Coefficients

Test statistic and decision

- The test statistic is

$$t = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{a_{jj}}} \quad \left(\text{or } F = t^2 = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2 a_{jj}} \right)$$

- Under the null hypothesis this statistic is t-distributed with $n - p - 1$ degrees of freedom (or F-distributed with $(1, n - p - 1)$ degrees of freedom)
- The null hypothesis is rejected if

$$|t| > t_{1-\alpha/2}(n - p - 1) \quad (\text{or } F > F_{1-\alpha}(1,\, n - p - 1))$$

- In contrast to the overall F-value, $F = t^2$ is said to be a *partial F-value*

Example: Testing Individual Regression Coefficients

- We compute

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} = \frac{3.75}{2} \cdot \frac{1}{1016} \begin{pmatrix} 1220 & -140 & -72 \\ -140 & 116 & -100 \\ -72 & -100 & 130 \end{pmatrix}$$

- and obtain

j	$\hat{\beta}_j$	a_{jj}	$t(2)$	$F(1,2)$
0	0.50	$\frac{1220}{1016} = 1.201$	0.333	0.111
1	1.00	$\frac{116}{1016} = 0.114$	2.161	4.670
2	-0.25	$\frac{130}{1016} = 0.128$	-0.510	0.260

- With $t_{0.975}(2) = 4.30$ or $F_{0.95}(1,2) = 18.51$ the null hypothesis cannot be rejected for any of the coefficients

Incremental F-Test

- The introduced tests can be conceived as special cases of a more general test procedure, which amounts to a comparison of models based on their likelihoods
- The incremental F-test compares two nested linear models
- Let M_1 denote the complex model and M_0 the simpler model resulting from restricting parameters, for example $\beta_j = 0$ for some j
- Definitions

SS_{reg0}, SS_{res0}	Explained and residual sum of squares of M_0
SS_{reg1}, SS_{res1}	Explained and residual sum of squares of M_1
q_0	Number of parameters in M_0
q_1	Number of parameters in M_1

Incremental F-Test

- With the statistic

$$
\begin{aligned}
F &= \frac{(SS_{reg1} - SS_{reg0})/(q_1 - q_0)}{SS_{res1}/(n - q_1)} = \frac{(SS_{res0} - SS_{res1})/(q_1 - q_0)}{SS_{res1}/(n - q_1)} \\
&= \frac{n - q_1}{q_1 - q_0} \cdot \frac{R_1^2 - R_0^2}{1 - R_1^2} \\
F &\sim F(q_1 - q_0, n - q_1) \text{ under } H_0
\end{aligned}
$$

  we test whether M_0 (and thus its restrictions) can be rejected in favor of M_1

- It can be shown that the test statistic can be derived from the general likelihood ratio test

$$
G^2 = 2 \log \frac{L_1}{L_0} = \frac{SS_{res0} - SS_{res1}}{\sigma^2} \overset{\hat{\sigma}^2 = \frac{q_1 - q_0}{n - q_1} SS_{res1}}{\longrightarrow} F = \frac{SS_{res0} - SS_{res1}}{\hat{\sigma}^2}
$$

  where L_0, L_1 are the likelihoods of the models M_0, M_1

Example: Incremental F-Test

- Let M_0 and M_1 be nested linear regression models defined by

$$\begin{aligned} M_0: &\quad Y_i = \beta_0 + \varepsilon_i \\ M_1: &\quad Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \end{aligned}$$

  with $\varepsilon_i \sim N(0, \sigma^2)$

- Because of

$$\sum_i (y_i - \bar{y})^2 = SS_{tot} = SS_{res0} \Rightarrow SS_{reg0} = 0$$

  the test statistic

$$F = \frac{(SS_{reg1} - 0)/(2-1)}{SS_{res1}/(n-2)} \sim F(1, n-2)$$

  reduces to the overall F-Test

Example: Cheddar Cheese

Model fitting

```
dat <- read.table("cheddar.txt", header = TRUE)

m1 <- lm(taste ~ Acetic + H2S + Lactic, dat)  # regression model
m0 <- lm(taste ~ 1, dat)                      # baseline model

coef(m1)                                      # OLS estimates
vcov(m1)
confint(m1)
```

Hypothesis testing

```
summary(m1)                                   # partial/overall F-tests
anova(m0, m1)                                 # incremental F-test
```

Example: Cheddar Cheese

Checking model assumptions

```
par(mfrow = c(1, 2))                    # split plotting window
qqnorm(resid(m1), main = "Observed")    # Q-Q plot
qqline(resid(m1))
qqnorm(y <- rnorm(30), main = "Simulated")
qqline(y)

par(mfrow = c(1, 1))
plot(resid(m1) ~ fitted(m1),            # residuals vs. fitted values
     main = "", xlab = "Predicted values", ylab = "Residuals")
abline(h = 0, lty = 2)

par(mfrow = c(2, 2))
plot(m1)                                # four diagnostic plots
```

Example: Cheddar Cheese

Influential observations

```
n <- nrow(dat)
p <- 3                                  # number of predictors
X <- model.matrix(m1)

hatvalues(m1)
H <- X %*% solve(t(X) %*% X) %*% t(X)   # hat matrix
h <- diag(H)

## Cook's distance
cooks.distance(m1)
s <- sigma(m1)                          # error standard deviation
resid(m1)^2 / ((p + 1) * s^2) * (h / (1 - h)^2)
```

General Linear Hypotheses

Objective and hypotheses

- The goal is to test whether linear combinations of the parameter vector $\boldsymbol{\beta}$ are distinct from a fixed vector $\boldsymbol{\xi}$
- This amounts to a simultaneous test of different hypotheses about the parameter vector
- It requires to specify a matrix that captures the hypotheses in the following way

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\xi}$$
$$H_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\xi}$$

General Linear Hypotheses

Examples for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$

- Testing the null hypothesis

$$H_0: \ \beta_1 = 0$$

- Here, we set $\mathbf{C} = (0\ 1\ 0\ 0)$ because

$$\begin{array}{lll} \mathbf{C} & \boldsymbol{\beta} & = \boldsymbol{\xi} \\ (0\ 1\ 0\ 0) & \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} & = 0 \end{array}$$

General Linear Hypotheses

Examples for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$

- Testing the null hypothesis

$$H_0 : \ \beta_2 = \beta_3 = 0$$

- Here, we set

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

because

$$\begin{array}{ccc} \mathbf{C} & \boldsymbol{\beta} & = & \boldsymbol{\xi} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} & \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} & = & \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{array}$$

General Linear Hypotheses

Examples for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$

- Testing the null hypothesis

$$H_0 : \ \beta_1 = \beta_2 \ \Leftrightarrow \ \beta_1 - \beta_2 = 0$$

- Here, we set $\mathbf{C} = (0\ 1\ -1\ 0)$ because

$$\begin{array}{ccc} \mathbf{C} & \boldsymbol{\beta} & = \ \boldsymbol{\xi} \\ (0\ 1\ -1\ 0) & \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} & = \ 0 \end{array}$$

General Linear Hypotheses

Test statistic and decision

- The test statistic

$$F = \frac{Q/r}{SS_{res}/(n-p-1)} = \frac{Q/r}{\hat{\sigma}^2}$$

is $F(r,\, n-p-1)$-distributed under the null hypothesis, where r is the rank of $\mathbf{C}$ (often the number of rows), and

$$Q = [\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}]' \, [\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1} \, [\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\xi}]$$

- The null hypothesis H_0 is rejected if

$$F > F_{1-\alpha}(r, n-p-1)$$

Example: General Linear Hypotheses

- For $\hat{\boldsymbol{\beta}} = (0.5,\ 1,\ -0.25)'$ we want to test the null hypothesis

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

- To derive the test statistic, we calculate for Q

$$\left[\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}0.5\\ 1\\ -0.25\end{pmatrix} - \begin{pmatrix}0\\ 0\end{pmatrix}\right]' \left[\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\frac{1}{1016}\begin{pmatrix}1220 & -140 & -72\\ -140 & 116 & -100\\ -72 & -100 & 130\end{pmatrix}\begin{pmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{pmatrix}\right]^{-1}$$
$$\left[\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\begin{pmatrix}0.5\\ 1\\ -0.25\end{pmatrix} - \begin{pmatrix}0\\ 0\end{pmatrix}\right]$$

$$= \frac{1016}{116 \cdot 130 - 100^2}\,(1\ -0.25)\begin{pmatrix}130 & 100\\ 100 & 116\end{pmatrix}\begin{pmatrix}1\\ -0.25\end{pmatrix} = 17.45$$

Example: General Linear Hypotheses

- With

$$Q = 17.45$$

and

$$\text{rk}(\mathbf{C}) = 2$$

we obtain the test statistic

$$F = \frac{17.45/2}{3.75/(5-2-1)} = 4.65$$

- As $F_{0.95}(2,2) = 19.00$ the null hypothesis is not rejected

Multicollinearity

Basics

- Multicollinearity describes the mutual linear dependency of variables in the context of multivariate procedures
- Multicollinearity occurs if predictor variables or their linear combinations show substantial intercorrelations
- Two vectors are called *collinear* if they fall on the same line (in case of a perfect correlation ± 1)

Multicollinearity

Diagnosis

- Different indicators have been developed which can provide evidence for multicollinearity
- We consider the variance inflation factor (VIF) defined by

$$VIF_j = \frac{1}{1 - R_j^2}$$

  where R_j^2 is the coefficient of determination of the regression of the jth predictor variable on the remaining predictors
- There is no generally accepted rule of thumb, but often a value of $VIF > 10$ is considered critical

Consequences of Multicollinearity

- Multicollinearity limits the achievable multiple correlation R, or the proportion of explained variability R^2
  - In the case of a regression of Y on two predictor variables X_1 and X_2, it can be shown that

$$R^2 = r^2_{yx_1} + r^2_{yx_2.x_1(s)}$$

with

r_{yx_1}	bivariate correlation between Y and X_1
$r_{yx_2.x_1(s)}$	semi-partial correlation between Y and X_2 adjusting for (removing) the influence of X_1 (on X_2 only)

Example: Consequences of Multicollinearity

- Intercorrelations

	Y	X_1	X_2
Y	1.000	0.894	0.640
X_1	0.894	1.000	0.814
X_2	0.640	0.814	1.000

- With $r_{yx_1} = 0.894$ we calculate

$$r_{yx_2.x_1(s)} = \frac{r_{yx_2} - r_{x_2x_1}\, r_{yx_1}}{\sqrt{1 - r^2_{x_2x_1}}} = \frac{0.640 - 0.814 \cdot 0.894}{\sqrt{1 - 0.814^2}} = -0.151$$

- We obtain $R^2 = r^2_{yx_1} + r^2_{yx_2.x_1(s)} = 0.894^2 + (-0.151)^2 = 0.822$
- When adjusting for X_1, X_2 only explains 2.3 % of the variability

Consequences of Multicollinearity

- Multicollinearity makes it difficult to evaluate the importance of a predictor
  - Due to interrelations with other predictors, the magnitude (and sign!) of a coefficient depends on which other predictors are in the model
- Multicollinearity compromises the estimation of regression coefficients
  - It increases the variance of the estimators
  - The regression equation becomes more unstable, meaning that results are highly sensitive to (even small) changes in the data or in the model equation
  - Although the matrix $\mathbf{X}'\mathbf{X}$ could still have full rank (even with high multicollinearity), we could face numerical problems calculating the inverse $(\mathbf{X}'\mathbf{X})^{-1}$
- Multicollinearity may lead to systematic biases when computing test statistics (i.e., t or F values)

Validation

Basics

- When evaluating a regression model, we should focus on its ability to predict unseen (new) replication data
- So the model should aim at a high *generalizability*, as opposed to goodness of fit captured by coefficients like R^2
- *Validation* describes the application of specific procedures to assess the generalizability of a regression model
- Validation methods include
  - Cross-validation procedures
  - Adjustments applied to the coefficient of determination

Validation

Cross-validation

- k-fold cross-validation
  - The sample is divided into k (preferably) equally sized subsamples
  - For each of these k subsamples, the prediction quality is determined (e. g., MSE) using the regression equation estimated from the remaining $k - 1$ subsamples
  - The k results are then averaged to produce a joint measure of prediction quality
- Leave-one-out (LOO) is the special case of an n-fold cross-validation
  - Every single observation, one at a time, is left out and predicted by a regression model fitted on the $n - 1$ remaining observations

Validation

Adjusting the coefficient of determination

- Quite a few formulas have been suggested to estimate the reduction of prediction quality when going from the considered sample to the population or to an independent (new) sample
- A commonly used formula (suggested by Wherry) is

$$\hat{\rho}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{SS_{res}/(n-p-1)}{SS_{tot}/(n-1)}$$

  where $\hat{\rho}$ is an estimator for the multiple correlation ρ in the population
- This formula tries to estimate the proportion of explained variability in the population, and not in an independent sample drawn from it

Validation

Determination of sample size

- For multiple regression, power analysis aims primarily at ensuring the generalizability of the results
- What is the minimal sample size such that the external prediction quality differs only slightly from the calculated coefficient of determination R^2?
- In order to ensure this, the ratio of the sample size n to the number of predictors p should be $n/p \approx 15$ at least

Categorical Predictors

Dummy coding

- In order to enter categorial (qualitative) variables into the regression equation, they have to be coded as *dummy variables*
- For coding a variable A with I categories, $I - 1$ dummy variables are needed
- For all focus categories $i = 2, \ldots, I$, we define

$$X_i^A = \begin{cases} 1 & \text{in case of category } i \\ 0 & \text{else} \end{cases}$$

- If a person/unit falls into the reference category $i = 1$ then all dummy variables X_i^A have the value 0
- Effects estimated for dummy variables are interpreted with respect to the reference category

Example: Categorical Predictors

Gender effect

- A marketing researcher assumes that the demand for a product (Y) depends not only on its prize (X_1) but also on the consumer's gender
- A dummy variable (X_2) is introduced that encodes gender (1 female, 0 male)
- The design matrix $\mathbf{X}$ then becomes

	x_1	x_2
1	x_{i1}	1

for female and

	x_1	x_2
1	x_{j1}	0

for male consumers

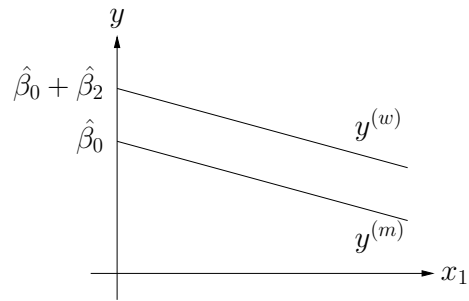
Example: Categorical Predictors

With $X_{i2} \in \{0, 1\}$, the general regression equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

becomes gender specific

$$Y_i^{(f)} = (\beta_0 + \beta_2) + \beta_1 X_{i1} + \varepsilon_i \qquad \text{female}$$
$$Y_j^{(m)} = \beta_0 + \beta_1 X_{j1} + \varepsilon_j \qquad \text{male}$$



- The β_2 parameter indicates a change in demand of female consumers compared to male consumers (reference)
- Using the $(0, 1)$-predictor for gender leads to a parallel shift of the lines, leaving the prize effect β_1 the same in both groups

Example: Categorical Predictors

Interaction effect

- If the prize influence on Y varies with the value of the dummy variable, an interaction effect is present
- We introduce a new dummy variable $X_3 = X_1 \cdot X_2$, so

$$X_3 = \begin{cases} X_1 \cdot 0 = 0 & \text{male consumer (reference)} \\ X_1 \cdot 1 = X_1 & \text{female consumer} \end{cases}$$

- The respective rows in the design matrix $\mathbf{X}$ then read

	x_1	x_2	x_3	
1	x_{i1}	1	x_{i1}	female
1	x_{j1}	0	0	male

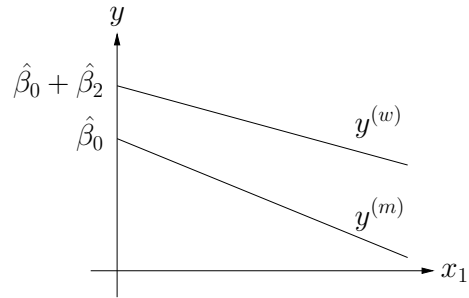
Example: Categorical Predictors

The general regression equation then is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

After inserting values for X_2 and X_3 it is

$$\begin{aligned} Y_i^{(f)} &= \beta_0 + \beta_1 X_{i1} + \beta_2 + \beta_3 X_{i1} + \varepsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_{i1} + \varepsilon_i \\ Y_j^{(m)} &= \beta_0 + \beta_1 X_{j1} + \varepsilon_j \end{aligned}$$



- The regression lines differ with respect to intercept (β_0 vs. $\beta_0 + \beta_2$) and slope (β_1 vs. $\beta_1 + \beta_3$)
- β_3 represents the prize effect for female consumers that comes on top of the male (reference) prize effect (group difference of effects)

Example: Cheddar Cheese

Linear hypothesis $\beta_1 = 0$ and $\beta_2 = \beta_3$

```
m1 <- lm(taste ~ Acetic + H2S + Lactic, dat)
X <- model.matrix(m1)

C <- matrix(c(0,1,0,0, 0,0,1,-1), 2, 4, byrow = TRUE)
diff <- C %*% coef(m1) - c(0, 0)
D <- solve(C %*% solve(t(X) %*% X) %*% t(C))
Q <- t(diff) %*% D %*% diff
r <- qr(C)$rank
F <- Q/r / sigma(m1)^2
qf(0.95, r, summary(m1)$df[2])      # F-quantile

car::linearHypothesis(m1, hypothesis.matrix = C, rhs = c(0, 0))
```

Equivalent incremental F-test

```
m2 <- lm(taste ~ I(H2S + Lactic), dat)
anova(m2, m1)
```

Example: Cheddar Cheese

Multicollinearity

```
cor(dat[, -1])        # correlation matrix
```

Variance inflation factor

```
1/(1 - summary(lm(Acetic ~ H2S + Lactic, dat))$r.squared)
1/(1 - summary(lm(H2S ~ Acetic + Lactic, dat))$r.squared)
1/(1 - summary(lm(Lactic ~ Acetic + H2S, dat))$r.squared)
```

Adjusted coefficient of determination

```
n <- nrow(dat)
p <- 3
1 - (n - 1)/(n - p - 1) * (1 - summary(m1)$r.squared)

summary(m1)$adj.r.squared
```

Simulation-Based Power Analysis

Background and code examples

- Wickelmaier, F. (2026). Simulating the power of statistical tests: a collection of R examples. arXiv. https://doi.org/10.48550/arXiv.2110.09836

Application scenario

- In a group of $n = 30$ patients, sampled from three different age groups (X_2), the strength of depressive symptoms (Y) will be measured. Each patient will receive a $0, 2, \ldots, 8$ mg dose (X_1) of medication. The goal is to detect either a linear reduction of symptoms of 2.5 points per mg dose in each age group, or a linear increase of 0.1 points per year of age in each medication group, or both, for $\alpha = 0.05$.
- What is the sample size necessary to detect this effect with 80% power?

Simulation-Based Power Analysis

- Model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad i = 1, \ldots, n$
- Hypothesis H_0: $\beta_1 = \beta_2 = 0$

```
n <- 30  # adjust sample size until target power is reached

dat <- expand.grid(x1 = rep(c(0, 2, 4, 6, 8), each = n/(5*3)),
                   x2 = c(30, 50, 80))

pval <- replicate(2000, {
    y <- 35 - 2.5*dat$x1 + 0.1*dat$x2 + rnorm(n, sd = 14)
    m0 <- lm(y ~ 1, dat)
    m1 <- lm(y ~ x1 + x2, dat)
    anova(m0, m1)$"Pr(>F)"[2]    # test H0
})
mean(pval < 0.05)                # power estimate
```

References

Draper, N. R. & Smith, H. (1998). Applied regression analysis. New York: Wiley.

Fahrmeir, L., Hamerle, A. & Tutz, G. (1996). Multivariate statistische Verfahren. 2. Auflage. Berlin: De Gruyter. (Kapitel 4)

Fox, J. (2008). Applied regression analysis and generalized linear models. Thousand Oaks, CA: Sage Publications.

Gelman, A. & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press.

Stevens, J. (2001). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. (Kapitel 3)