Lecture 3

Multiple linear regression

Let Y be a dependent variable (a response),

$X_1,\ldots, X_k$ be independent variables (factors).

Assume that we have n observations of a multivariate random variable

$(Y, X_1,\ldots, X_k)$ → $Y = f(X_1,\ldots, X_k) + \varepsilon$

A linear multiple regression model has the following form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon.$$

Here we get the theoretical regression equation

$$\overline{y}_x = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

where $\beta_1,\ldots,\beta_k$ are coefficients of multiple regression.

Additional Gauss-Markov assumption for multiply regression

!!! Variables $X_1, \ldots, X_k$ are pairwise not correlated.

The linear model in matrix form is:

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

$y = a + bx$

$y_i = a + bx_i$

x^0

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

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \ldots \beta_k x_{ik}$

where

x_1

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$n \times k+1$

Identification of a regression model (parameter estimation)

The theoretical regression equation is $\bar{\mathbf{y}}_{\mathbf{X}} = \mathbf{X}\boldsymbol{\beta}.$

The empirical regression equation is $\hat{\mathbf{y}}_{\mathbf{X}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$

$$\underbrace{A^{-1} = \frac{1}{|A|} \cdot \tilde{A}}$$

Here $\hat{\boldsymbol{\beta}}$ is the OLS-estimator, i.e. it is derived from the condition

$$\sum_{i=1}^{n} (y_i - \hat{y}_{X_i})^2 \to \min \qquad (*)$$

$X_{n \times (k+1)}$

$X^T_{(k+1) \times n}$

$$\underbrace{X^T}_{(k+1)\times n} \; X_{\cancel{n\times}(k+1)} =$$

$$(\quad)_{(k+1)\times(k+1)}$$

The solution of minimization problem (*) is the following vector

$A_{k \times k}$

$|A_{k \times k}| \neq 0$

$\exists A^{-1}$

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

$!\ |X^T X| \neq 0$

This takes place if $rang(\mathbf{X}) = k + 1.$

rank $= r(X \quad)$

The interpretation of a regression coefficient: it provides an estimated effect of a unit change of an independent variable given the other variables being fixed.

Coefficients of correlation and determination

The coefficient of determination: $R^2 = 1 - \frac{ESS}{TSS}$ R-square

The multiple correlation coefficient: $\hat{\rho}_{yx_1 \ldots x_k} = \sqrt{1 - \frac{ESS}{TSS}}$

The adjusted coefficient of determination: $R^2{}_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$

Adjusted R-square

Adjusted R-squared: You should choose the models that have higher adjusted R-squared value. These statistic is designed to avoid a key problem with regular **R-squared** that increases every time you add a predicted variable and can trick you into specifying an overly complex model.

It contains a change for the number of independent variables and is widely used for selection reasonable independent variables into the model.

Multicollinearity problem

In regression analysis, we look at correlations between one or more factors and a response.

With regression, as with so many things in life, there comes a point where adding *more* is not better.

- Multicollinearity is a phenomenon in which two or more factors in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy.
- Multicollinearity occurs when your model includes multiple factors that are statistically correlated not just to your response variable, but also to each other.
- If nominally "different" measures actually quantify the same phenomenon then they are redundant. In other words, it results when you have factors that are redundant.

We differentiate between:

- *perfect collinearity*, which means that there is a linear functional relationship between factors (correlation coefficient =1),
- *partial collinearity or multicollinearity*, for which there is a high correlation between the factors (more than 0,7; or less than -0,7).

Multicollinearity problem

In case of perfect multicollinearity, there is a violation one of the Gauss-Markov assumptions: independence of the explanatory variables.

- The ordinary least squares estimates involve inverting the matrix $\mathbf{X}^T\mathbf{X}$ However, in this case, the design matrix $\mathbf{X}$ is singular. So, we have

$$\det(\mathbf{X}^T\mathbf{X}) = 0.$$

- Under these circumstances, the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}$ does not exist.

⇒ It becomes impossible to identify the model.

- When partial multicollinearity occurs we get unstable OLS-estimates of the coefficients for a regression model.

⇒ *Multicollinearity increases the standard errors of the coefficients, making them unstable.*

⇒ *By overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they might be significant.*

⇒ *Small changes to the input data can lead to large changes in the model, even resulting in changes of sign of parameter estimates.*

It makes the coefficients to be not interpreted.

Warning Signs of Multicollinearity

How can you know if there is multicollinearity in your regression model?

Here are some things to watch for:

- A regression coefficient is not significant even though that variable is highly correlated with *Y*.
- When you add or delete an *X* variable, the regression coefficients change dramatically.
- You see a negative regression coefficient when your response should increase along with *X*.
- You see a positive regression coefficient when the response should decrease as *X* increases.
- Your *X* variables have high pairwise correlations.

Warning Signs of Multicollinearity

The ways to detect

1) To construct *the matrix of pairwise correlation coefficients* and check its determinant

$$\det(\mathbf{X}^T\mathbf{X}) \approx 0.$$

The closeness to 0 the determinant of the matrix $\mathbf{X}^T\mathbf{X}$ means a high degree of multicollinearity and, consequently, unreliable results of regression analysis.

The proximity this determinant to 1 indicates a good model specification.

2) To construct a linear regression model for each factor X_j relatively to all other factors and calculate for these models coefficients of determination:

$$R^2_{x_1\|x}, R^2_{x_2\|x}, \dots, R^2_{x_k\|x}.$$

Into the model there should be selected the factors with minimum value $R^2_{x_j\|x}$.

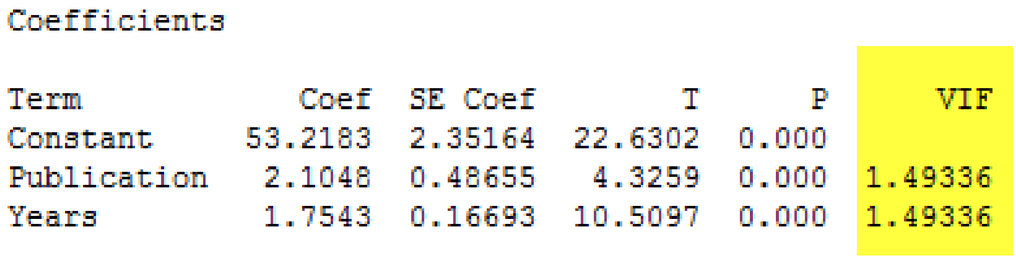
3) One more way to measure multicollinearity is *the variance inflation factor* (VIF), which assesses how much the variance of an estimated regression coefficient increases if your factors are correlated.

If no factors are correlated, the VIFs will all be 1.

4) To perturb the data. Multicollinearity can be detected by adding random noise to the data and re-running the regression many times and seeing how much the coefficients change.

Example

Here is an example involving some data looking at the relationship between *Researcher salary, Publications, and Years of employment*:



```
Coefficients

Term            Coef  SE Coef        T      P      VIF
Constant     53.2183  2.35164  22.6302  0.000
Publication   2.1048  0.48655   4.3259  0.000  1.49336
Years         1.7543  0.16693  10.5097  0.000  1.49336
```

$$tolerance_{x_j} = 1 - R^2_{x_j \| x}$$

$$VIF_{x_j} = \frac{1}{tolerance_{x_j}}$$

- If the VIF is equal to 1 there is no multicollinearity among factors.
- If the VIF is greater than 1, the predictors may be moderately correlated.
- The output above shows that the VIF for the Publication and Years factors are about 1.5, which indicates some correlation, but not enough to be overly concerned about.
- A VIF between 5 and 10 indicates high correlation that may be problematic.
- And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

*Jim Frost «The Minitab Blog»
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-tutorial-and-examples

Choosing the correct regression model is as much a science as it is an art.

It starts when we want to mathematically describe the relationship between some factors and a dependent variable. We investigate typically measures many variables but includes only some of them in the model. Along the way, we have to consider many possible models.

Too few factors: An underspecified model tends to produce biased estimates.
Too many factors: An overspecified model tends to have less precise estimates.
Just right: A model with the correct terms has no bias and the most precise estimates.

We review some common statistical methods for selecting models, complications you may face, and provide some practical advice for choosing the best regression model.

Stepwise regression

Stepwise regression selects a model by automatically adding or removing individual factors, a step at a time, based on their statistical significance. The end result of this process is a single regression model, which makes it nice and simple.

Best Subsets Regression

Best Subsets compares all possible models using a specified set of factors, and displays the best-fitting (according to some chosen criteria) models that contain one factor, two factors, and so on. The end result is a number of models and their summary statistics. It is up to you to compare and choose one.

Selection factors for a model

Methods:

- Stepwise regression (the inclusion method, the exclusion method, and both)
- Best subsets regression

These are two automated procedures that can identify useful factors during the exploratory stages of model building.

Useful statistics:

1. Adjusted R-squared increases only if the new term improves the model more than would be expected by chance and it can also decrease with poor quality factors.
2. P-values for the factors: In regression, low p-values indicate terms that are statistically significant. “Reducing the model” refers to the practice of including all candidate factors in the model, and then systematically removing the term with the highest p-value one-by-one until you are left with only significant factors.
3. Mallows’ Cp: it is a statistic specifically designed to help you manage the tradeoff between precision and bias.

Example

Imagine that you have a small delivery company that provides delivery services (deliver letters, packages and etcetera). You need to calculate an approximate travel time for different variables influenced on the total time.

To conduct your analysis you take a random sample of 10 past trips and record the following information:

- X1 – total miles traveled (in miles)
- X2 – number of deliveries (in units)
- X3 – the daily gas price (in $)
- Y – total travel time (in hours)

X1 milesTraveled	X2 numDeliveries	X3 gasPrice	Y travelTime (hrs)
89	4	3,84	7
66	1	3,19	5,4
78	3	3,78	6,6
111	6	3,89	7,4
44	1	3,57	4,8
77	3	3,57	6,4
80	3	3,03	7
66	2	3,51	5,6
109	5	3,54	7,3
76	3	3,25	6,4

Example (results)

Conducting regression analysis we have

Model No	X1	X2	X3	F	p-value	SE	R-Sq (adj)	R-Sq	VIF
1	x			49,77	<0,001	0,34	84,42%	86,15%	1
2		x		41,96	<0,001	0,37	81,99%	83,99%	1
3			x	0,62	0,455	0,89	0,00%	7,14%	1
4	x	x		23,72	0,001	0,35	83,47%	87,14%	11,59
5	x		x	22,63	0,001	0,35	82,78%	86,61%	1,14
6		x	x	27,63	<0,001	0,32	85,55%	88,76%	1,33
7	x	x	x	16,99	0,002	0,34	84,20%	89,50%	17,35

How to identify the most important factors in regression models

- Regular regression coefficients describe the relationship between each independent variable and the response.

- The coefficient value represents the mean change in the response given a one-unit increase in the factor.

Larger coefficients do not necessarily identify more important factors (!)

- Regular regression coefficients use different scales and you can not compare them directly.

- However, if you standardize the regression coefficients so they are based on the same scale, you can compare them.

Standardized coefficients of linear multiple regression

Regression equation is

$$y_x = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Subtracting the mean, then dividing by the standard deviation we get new variables called *standardized variables*

$$t_y = \frac{y - \bar{y}}{\sigma_y}, \qquad t_{x_j} = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}$$

Standardized regression equation has the form

$$\bar{t}_y = b_1 t_{x_1} + \ldots + b_k t_{x_k},$$

where b_j is *standardized coefficients.*

This coding puts the different factors on the same scale and allows you to compare their coefficients directly (!)

Relationship between regression and standardized coefficients:

$$\beta_j = b_j \frac{\sigma_y}{\sigma_{x_j}}, \quad j = 1, \ldots, k.$$

Importance of individual factors in regression models

These coefficients allow us to range the factors in the order of their influence on the dependent variable.

Standardized regression coefficients:

$$b_j = \beta_j \frac{\sigma_{x_j}}{\sigma_y}, \qquad j = 1, \dots, k.$$

A standardized coefficient represents the mean change in the response Y given a one standard deviation change in the factor X_j with the rest factors being fixed.

Look for the factor with the largest absolute value for the standardized coefficient.

Elasticity coefficients (partial and average):

- Elasticity is independent of units and thus simplifies data analysis.
- A high value of elasticity means a strong influence of an independent variable on the dependent variable.

Elasticity is the measurement of how responsive the dependent variable is to a change in a factor.

$$Э_{yx_j} = \frac{\partial y}{\partial x_j} \frac{x_j}{y}, \qquad \overline{Э}_{yx_j} = \beta_j \frac{\overline{x}_j}{\overline{y}}, \qquad j = 1, \dots, k.$$

An average coefficient of elasticity is interpreted in this way: with increasing factor X_j at 1% of the average level the dependent variable increases by $\overline{Э}_{yx_j}$% from its average level when the other factors are held constant.

These statistics might not agree because the manner they defines "most important" is a bit different.

Durbin-Watson criterion

Durbin-Watson statistic

$$d = \frac{\sum_{i=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{i=1}^{n} \varepsilon_t^{\ 2}}.$$

$$\boxed{d \approx 2 \cdot (1 - r_1^{\varepsilon}).}$$

The first order coefficient of correlation between residuals:

$$r_1^{\varepsilon} = \frac{\sum_{i=2}^{n} (\varepsilon_t - \bar{\varepsilon}_1)(\varepsilon_{t-1} - \bar{\varepsilon}_2)}{\sqrt{\sum_{i=2}^{n} (\varepsilon_t - \bar{\varepsilon}_1)^2 \cdot \sum_{i=2}^{n} (\varepsilon_{t-1} - \bar{\varepsilon}_2)^2}},$$

where

$$\bar{\varepsilon}_1 = \frac{\sum_{i=2}^{n} \varepsilon_t}{n-1}; \quad \bar{\varepsilon}_2 = \frac{\sum_{i=2}^{n} \varepsilon_{t-1}}{n-1}.$$

Есть положительная автокорреляция остатков. H_0 отклоняется. С вероятностью $P = (1–\alpha)$ принимается H_1.	Зона неопределенности	Нет оснований отклонять H_0 (автокорреляция остатков отсутствует)	Зона неопределенности	Есть отрицательная автокорреляция остатков. H_0 отклоняется. С вероятностью $P = (1-\alpha)$ принимается H_1*

0 d_L d_U 2 4-d_U 4-d_L
4

Application limits:

1. The test is not designed for detection other types of autocorrelation (higher than the first order).
2. In the model the constant term must be presented.
3. Lagged variables are not permitted to include in the model.

Heteroscedasticity

Methods of detection

- Graphic (visual) analysis
- Using statistical tests:
  - Spearman's rank correlation test
  - Goldfeld-Quandt test
  - Glazer test
  - Breusch-Pagan test
  - White test, and others.

Remedies

1) Transform variables (log transformation);
2) Change the model specification;
3) Add new data and run the new model identification;
4) Use a weighted least squares;
5) Use the generalized least squares (GLS).

Autocorrelation in the residuals

Methods of detection

- Graphic (visual) analysis
- Series method
- Using statistical tests:
  - Durbin-Watson test
  - Breusch-Godfrey test
  - Q-test Ljung-Box test
  - And others.

Remedies

1) Apply the first order autoregressive scheme - AR(1).
2) Include into the model time variable *t* or lagged variables *Yt-1, Yt-2*, etc.
3) Transform the data by taking first differences.
4) Use GLS.