Multiple variables linear regression

Statistics

Anton Afanasev

Higher School of Economics

DSBA 211
March 18, 2023

Seminar Overview

Quiz

A 95% confidence interval for a regression slope was calculated on the basis of 1000 observations:

$$(\beta_1)_{95\%} \in (0.11, 0.65).$$

Calculate the p-value for the null hypothesis that Y does not increase with X.

Problem 1

Suppose that a random sample of 4 families has the following annual incomes and savings:

Family	Income X (Thousands of \$)	Savings S (Thousands of \$)
A	22	2.0
B	18	2.0
C	17	1.6
D	27	3.2

1. Estimate the population regression $S = \beta_0 + \beta_1 X$.
2. Construct a 95% confidence interval for the slope β_1.
3. Graph the four points and the fitted line, and then graph as well as you can the acceptable slopes given by the confidence interval in part 2.
4. Which of the following hypotheses are rejected by the data at the 5% level?

$$\beta_1 = 0? \qquad \beta_1 = 0.05? \qquad \beta_1 = 0.10? \qquad \beta_1 = 0.50?$$

Multiple variables linear regression model

- Linear regression model of n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

where $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix}^\top$ – vector of p regressors,

y – regressand,

β_0 – intercept of true regression hyperplane,

$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_p \end{pmatrix}^\top$ – vector of regressor effects,

ε – disturbance term with $\mathsf{E}(\varepsilon) = 0$ and $\mathsf{V}(\varepsilon) = \sigma^2 > 0$.

- True regression hyperplane:

$$\mathsf{E}(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}.$$

- Assumptions (to satisfy Gauss-Markov theorem):
  1. values of regressors vector $\mathbf{x}_i$ are constants,
  2. noise instants are uncorrelated: $\mathsf{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j$.

Sample regression hyperplane

- Under assumptions from above y_i are uncorrelated random variables with

$$\mathsf{E}(y_i) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i \quad \text{and} \quad \mathsf{V}(y_i) = \sigma^2.$$

- Sample regression hyperplane:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\top \mathbf{x}.$$

- Values of regressand on the sample hyperplane, corresponding to vectors $\mathbf{x}_i$:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i$$

- Estimate of ε_i is a difference between observation y_i and estimate $\widehat{y}_i$, denoted as e_i:

$$e_i = \widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i.$$

OLS estimates

- Quadratic minimization of $\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2$:

$$\left(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}\right) = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2.$$

- In order to find explicit results for $\left(\widehat{\beta}_0, \widehat{\boldsymbol{\beta}}\right)$ let's denote a matrix of regressors as follows:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Result:

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\boldsymbol{\beta}} \end{pmatrix} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y},$$

where $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^\top$ – vector of regressands.

Moments of $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$

- Expected values:

$$\mathsf{E}\left(\widehat{\beta}_0\right) = \beta_0, \qquad \mathsf{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}.$$

- Covariance matrix:

$$\mathsf{Cov}\begin{pmatrix}\widehat{\beta}_0\\ \widehat{\boldsymbol{\beta}}\end{pmatrix} = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.$$

- Standard errors:

$$\text{S.E.}\left(\widehat{\beta}_j\right) = \sigma\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}},$$

where $(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}$ is the j^{th} diagonal element of matrix $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

- Equation on $\widehat{\beta}_0$ via $\widehat{\boldsymbol{\beta}}$ (as in simple linear regression):

$$\widehat{\beta}_0 = \overline{y} - \widehat{\boldsymbol{\beta}}^\top\overline{\mathbf{x}},$$

where $\overline{\mathbf{x}} = \begin{pmatrix}\overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p\end{pmatrix}^\top$.

Estimation of variance

- Model of multiple variables linear regression:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad \mathsf{E}(\varepsilon_i) = 0, \mathsf{V}(\varepsilon_i) = \sigma^2.$$

- Variation, created by ε_i (residual, error):

$$RSS = SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i\right)^2$$

with $n - p - 1$ degrees of freedom.

- Point estimate of σ^2:

$$\widehat{\sigma^2} = \frac{RSS}{n - p - 1} = MSE.$$

Normal regression

- Now assume $\varepsilon_1, \ldots, \varepsilon_n \sim$ i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$.
- Subsequently $y_i \sim \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i, \sigma^2\right)$ and they are independent.
- Linear combination of normal variables is normal:

$$\widehat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}_{jj}\right).$$

- Pivot functions to estimate β_j:

$$\frac{\widehat{\beta}_j - \beta_j}{\text{S.E.}\left(\widehat{\beta}_j\right)} \sim \mathcal{N}(0, 1).$$

- Can't be used when σ^2 is unknown (almost always).

Estimated standard errors of $\widehat{\beta}_j$

- Let's replace standard errors with estimated standard errors, where we replace σ^2 with its point estimate – MSE:

$$\text{E.S.E.}\left(\widehat{\beta}_j\right) = \sqrt{MSE \cdot (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}.$$

- According to Fisher's lemma:

$$\frac{(n-p-1)\cdot MSE}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p-1}.$$

- Pivot functions then:

$$\frac{\widehat{\beta}_j - \beta_j}{\text{E.S.E.}\left(\widehat{\beta}_j\right)} \sim t_{n-p-1}.$$

- MSE is independent of $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$.

Intervals estimation for β_j

- $(1-\alpha)\%$ confidence intervals for β_j:

$$(\beta_j)_{1-\alpha} \in \widehat{\beta}_j \pm t_{n-p-1;\, \alpha/2} \cdot \text{E.S.E.}\left(\widehat{\beta}_j\right).$$

- Test statistic for $\mathrm{H}_0 : \beta_j = b_j$:

$$T_{n-p-1}\Big|_{\mathrm{H}_0} = \frac{\widehat{\beta}_j - b_j}{\text{E.S.E.}\left(\widehat{\beta}_j\right)}.$$

Sources of variance in multiple variables linear regression

- Model of multiple variables linear regression:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad \mathsf{E}(\varepsilon_i) = 0, \mathsf{V}(\varepsilon_i) = \sigma^2.$$

  Regressand y can change because of:

$$\begin{aligned} \varepsilon_i &- \text{error/noise}, \\ \boldsymbol{\beta}^\top \mathbf{x}_i &- \text{change in regressors vector } \mathbf{x}. \end{aligned}$$

- We can compare variances, created by both sources, to find a significant evidence of presence of dependency between $\mathbf{x}$ and y.

$$\text{Total SS} = \underbrace{\text{Regression SS}}_{\boldsymbol{\beta}^\top \mathbf{x}_i} + \underbrace{\text{Residual SS}}_{\varepsilon_i}$$

Variation, produced by regression

- Sample regression hyperplane with OLS estimate of β_0:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i,$$
$$\widehat{y}_i = \overline{y} - \widehat{\boldsymbol{\beta}}^\top \overline{\mathbf{x}} + \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i.$$

- Deviation of regressand estimate from the mean is in direct ratio to the deviation of regressors vector:

$$\widehat{y}_i - \overline{y} = \widehat{\boldsymbol{\beta}}^\top (\mathbf{x}_i - \overline{\mathbf{x}}).$$

This difference will produce regression variation:

$$(y_i - \overline{y}) = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i),$$
$$(y_i - \overline{y}) = \widehat{\boldsymbol{\beta}}^\top (\mathbf{x}_i - \overline{\mathbf{x}}) + (y_i - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i).$$

ANOVA decomposition

- Total variation decomposition:

$$
\begin{array}{ccccc}
\text{Total SS} & = & \text{Regression SS} & + & \text{Residual SS} \\
TSS & = & ESS & + & RSS \\
SST & = & SSR & + & SSE \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 & = & \sum_{i=1}^{n}\left[\widehat{\boldsymbol{\beta}}^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right]^2 & + & \sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\boldsymbol{\beta}}^\top \mathbf{x}_i)^2
\end{array}
$$

with degrees of freedom

$$
\begin{array}{ccccc}
n-1 & = & p & + & n-p-1
\end{array}
$$

Pivot function

- Under assumptions of normal regression ($\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$) and truth of null hypothesis of ANOVA in multiple variables linear regression
$$\mathrm{H}_0 : \beta_1 = \ldots = \beta_p = 0,$$
the following is true:
$$\begin{aligned} SST/\sigma^2 &\sim \chi^2_{n-1}, \\ SSR/\sigma^2 &\sim \chi^2_{p}, \\ SSE/\sigma^2 &\sim \chi^2_{n-p-1}. \end{aligned}$$

- Pivot function:
$$F\Big|_{\mathrm{H}_0} = \frac{SSR/p}{SSE/(n-p-1)} \sim F_{p;\, n-p-1}.$$
H_0 is rejected in favor of H_1 : not H_0 if $F > F_{p;\, n-p-1;\, \alpha}$.

Adjusted coefficient of determination

- R^2 automatically increases when extra explanatory variables are added to the model.
- To penalize for the excess number of regressors which do not add to the explanatory power of the regression, adjusted coefficient of determination is introduced:

$$\overline{R}^2 = 1 - \frac{SSE/\mathrm{DF}_{\mathrm{error}}}{SST/\mathrm{DF}_{\mathrm{total}}}.$$

- In a case of multiple variables linear regression:

$$\overline{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1}.$$

- $\overline{R}^2$ can be negative for poorly fitting models.

Problem 2

HSE student Ivan is given a set of 20 observations (x_i, y_i). Teacher asks him to construct a simple regression line on the basis of those observations and check its goodness-of-fit. But Ivan is very observant, and he finds that in addition to the linear trend, the set has periodic oscillation trend.

x	-6.99	7.07	9.52	1.86	-5.62	-7.08	6.18	0.32	-7.27	-4.31
y	-7.01	9.23	9.97	3.82	-3.78	-8.63	5.21	0.36	-8.38	-3.93

x	1.14	-7.12	-8.24	8.26	8.94	-8.43	7.32	1.22	3.03	6.52
y	3.09	-7.48	-11.06	8.24	10.04	-11.82	12.00	3.71	3.21	7.70

Problem 2

1. Ivan makes the assumption – observations have following form:

$$y_i = \alpha x_i + \beta \sin x_i + \varepsilon_i,$$

   where ε_i is the Gaussian noise with zero mean.
   Find OLS regression curve, which satisfies this form of dependency between y and x.

2. After receiving results based on his own insight, Ivan decides to check the fit of the simple regression line, which is known to be derived from

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

   Find the corresponding regression line.

3. Compare the goodness-of-fit of 2 models in terms of coefficient of determination. Which one is better?

Problem 2

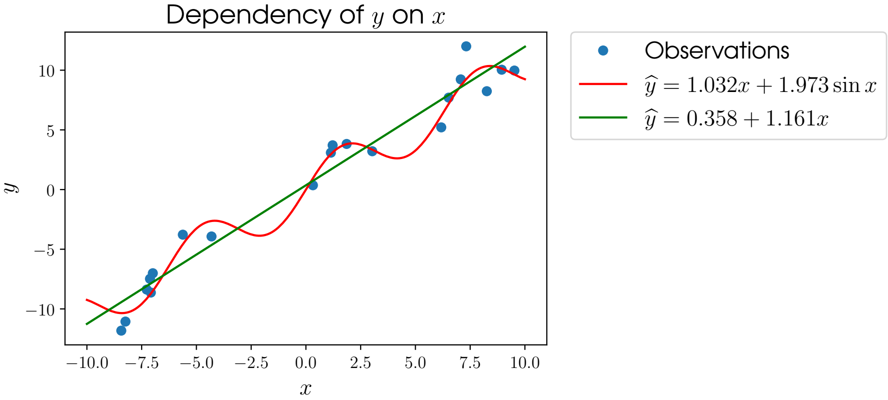

Figure: Comparison of models.

Problem 3

Suppose you are given a set of n observations:

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \varepsilon_i, \quad i = \overline{1, n},$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ is a vector of m unknown parameters, y_i is an observation of dependent variable, $\mathbf{x}_i \in \mathbb{R}^m$ is a vector of observations of m regressors, and ε_i is a noise. All ε_i are i.i.d. random variables with $\mathsf{E}(\varepsilon_i) = 0$ and $\mathsf{V}(\varepsilon_i) > 0$.
Find the MLE of vector $\boldsymbol{\beta}$ if:

1. $\varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$,
2. $\varepsilon_i \sim \text{Laplace}\,(0, a)$,
3. $\varepsilon_i \sim \mathcal{U}\left(-a, a\right)$.

Problem 4

Let us have classical regression model $y_t = \beta x_t + \varepsilon_t$, and $x_t > 0, y_t > 0$ for all $t = \overline{1, n}$.

Consider estimators $\beta^* = \dfrac{\overline{y}}{\overline{x}}$ and $\widetilde{\beta} = \dfrac{1}{n} \sum\limits_{t=1}^{n} \dfrac{y_t}{x_t}$.

1. Derive the LS estimator $\widehat{\beta}$ and check if it's unbiased.
2. Are estimators β^* and $\widetilde{\beta}$ unbiased?
3. Find variances of all three estimators.
4. Compare their variances. Which estimator is the best in terms of MSE?

Useful inequalities

Cauchy's inequality in Euclidean space $\mathbb{R}^n$

$$\forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n : \quad \left(\sum_{i=1}^{n} u_i v_i \right)^2 \leq \left(\sum_{i=1}^{n} u_i^2 \right) \left(\sum_{i=1}^{n} v_i^2 \right)$$

where u_i and v_i are components of $\mathbf{u}$ and $\mathbf{v}$ respectively.

Hölder's inequality

$$\forall m, n \in \mathbb{N} : \quad \left(\sum_{i=1}^{n} \alpha_i \beta_i \cdots \omega_i \right)^m \leq \left(\sum_{i=1}^{n} \alpha_i^m \right) \left(\sum_{i=1}^{n} \beta_i^m \right) \cdots \left(\sum_{i=1}^{n} \omega_i^m \right)$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta}, \ldots, \boldsymbol{\omega} \in \mathbb{R}^n$ are m vectors with non-negative components $\alpha_i, \beta_i, \ldots, \omega_i$ respectively.

Look at the time!