SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 3 – Multiple linear regression (MPV 3, Iz 5.2)

February 14, 2022

Todays lecture

- Multiple linear regression models
- Matrix notation for MLRM
- Least squares for MLRM
- The hat matrix
- Properties of the LS estimators
- Estimation of σ^2
- Random regression variables

Multiple linear regression models

In many situations we need a more complicated model. Some examples of such models are given by, e.g.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

or more generally

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

These are said to be *multiple linear regression models* because they are linear in the *regression coefficients* β_0, $\beta_1, \ldots,$ β_k. x_1, $x_2, \ldots,$ x_k are called *regressors* or *prediction variables*.

Multiple linear regression models

More complex models, such as e.g.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \varepsilon$$

can be treated by methods for multiple linear regression models by letting $x_2 \coloneqq x_1^2$ and $x_3 \coloneqq x_1^3$. Similarly,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 e^{x_3} + \varepsilon$$

can be treated by methods for multiple linear regression models by letting $x_3 \coloneqq x_1 x_2$ and $x_4 \coloneqq e^{x_3}$.

Matrix notation

Let $\mathbf{x}_i := (1, x_{i1}, x_{i2}, \ldots, x_{ik})$ denote the ith observation of $(1, x_1, x_2, \ldots, x_k)$. Given n observations, we define

$$\underbrace{X := \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ 1 & x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}}_{\text{The model matrix}} \quad \boldsymbol{\beta} := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\varepsilon} := \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Using this notation, the multiple linear regression model can be written as

$$\underbrace{\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}}_{\text{the regression function}}$$

Least squares

As in the 1-dimensional case, we want to choose $\beta_0,\ \beta_1, \ldots,\ \beta_k$ which minimizes the least-squares/loss function

$$\underbrace{S(\beta_1, \beta_2, \ldots,\ \beta_k)}_{\text{the loss function}} \coloneqq \sum \varepsilon_i^2 = \sum \big(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})\big)^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2.$$

The least-squares normal equations

The equations $\frac{dS}{\beta_1} = 0,\ \frac{dS}{\beta_2} = 0,\ \ldots,\ \frac{dS}{\beta_k} = 0$ are called the least-squares normal equations, and their solution the *least squares estimators* $\hat{\beta}_0,\ \hat{\beta}_1, \ldots,\ \hat{\beta}_k$.

Least squares

We have

$$
\begin{aligned}
S(\boldsymbol{\beta}) = \sum \varepsilon_i^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 &= (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X\boldsymbol{\beta} + (X\boldsymbol{\beta})^T X\boldsymbol{\beta} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}.
\end{aligned}
$$

Consequently, the least squares solution $\hat{\boldsymbol{\beta}}$ must satisfy

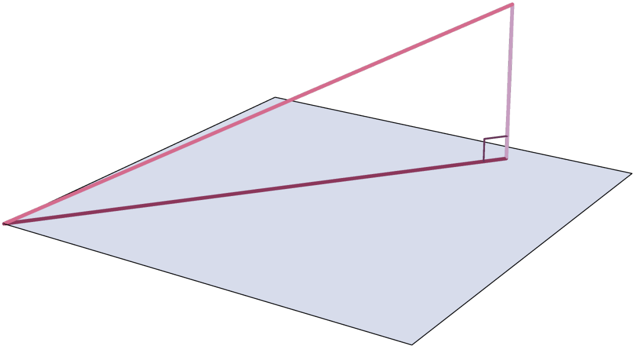
$$
\underbrace{0 = \left.\frac{dS}{d\boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = -2X^T\mathbf{y} + 2X^T X\hat{\boldsymbol{\beta}}}_{\text{least-squares normal equations}}
$$

Solving for $\hat{\boldsymbol{\beta}}$, we obtain

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

From linear algebra, we know that the matrix $X^T X$ in invertible exactly if the columns of $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$ are linearly independent.

Geometric interpretation



Example

```
pairs(df00[,3:13],col = "#703457", pch = 16, lwd=0, cex=.7,
    xaxt='n', yaxt='n')
```

Example

```r
df00.model <- lm(people_fully_vaccinated_per_hundred~gdp_per
    _capita, data = df00)
summary(df00.model)
```

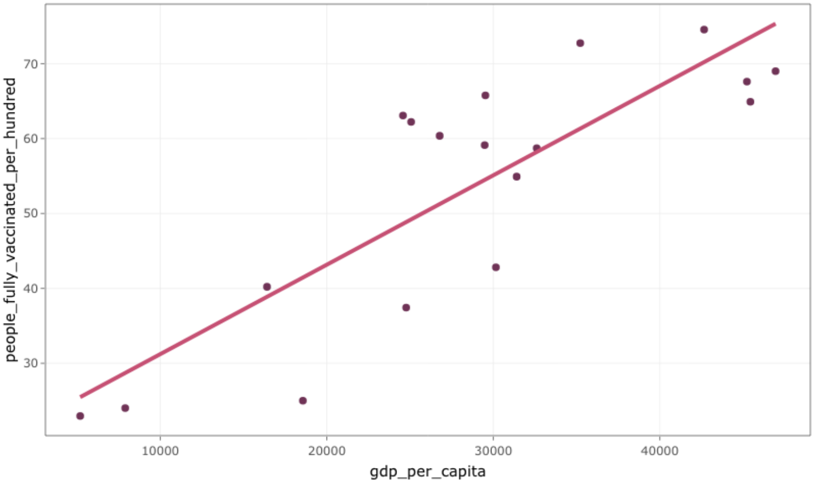
```
Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita, data = df00)

Residuals:
    Min      1Q  Median      3Q     Max
-16.428  -6.176  -0.675   7.997  14.445

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.929e+01  6.075e+00   3.175  0.00588 **
gdp_per_capita 1.194e-03  1.957e-04   6.100 1.53e-05 ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.665 on 16 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared: 0.6993,  Adjusted R-squared:  0.6805
F-statistic: 37.21 on 1 and 16 DF,  p-value: 1.534e-05
```

Example



people_fully_vaccinated_per_hundred
70
60
50
40
30
10000
20000
30000
40000
gdp_per_capita

Example

```r
df00.model2 <- lm(people_fully_vaccinated_per_hundred~gdp_
    per_capita+hospital_beds_per_thousand, data = df00)
summary(df00.model2)
```

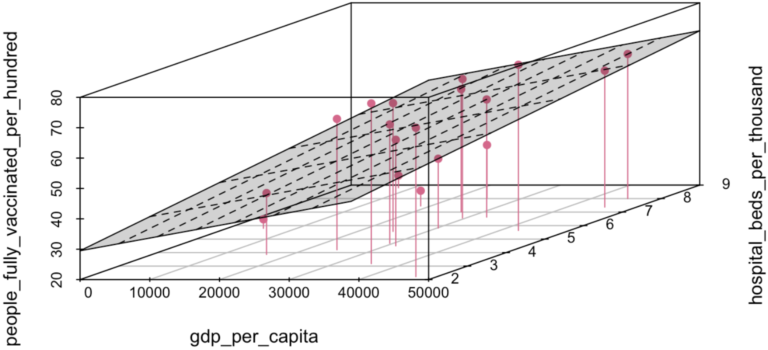
```
Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita + hospital_beds_per_thousand, data = df00)

Residuals:
     Min       1Q   Median       3Q      Max
-13.7639  -4.4811   0.0485   5.8690  12.7837

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    33.7305585 10.3210623   3.268  0.00519 **
gdp_per_capita  0.0011229  0.0001901   5.908 2.88e-05 ***
hospital_beds_ -2.1185866  1.2567801  -1.686  0.11253
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.153 on 15 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared: 0.7472,  Adjusted R-squared:  0.7135
F-statistic: 22.17 on 2 and 15 DF,  p-value: 3.318e-05
```

Example



people_fully_vaccinated_per_hundred
20
30
40
50
60
70
80
0
10000
20000
30000
40000
50000
gdp_per_capita
2
3
4
5
6
7
8
9
hospital_beds_per_thousand

Example

```
df00.model3 <- lm(people_fully_vaccinated_per_hundred~gdp_
    per_capita+I(gdp_per_capita^2), data = df00)
summary(df00.model3)
```

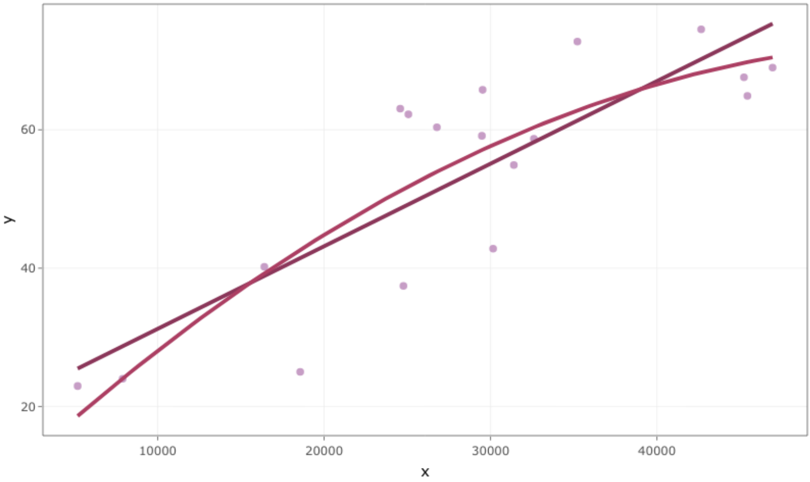
```
Call:
lm(formula = people_fully_vaccinated_per_hundred ~ gdp_per_
    capita + I(gdp_per_capita^2), data = df00)

Residuals:
     Min       1Q   Median       3Q      Max
-17.6333  -3.6830   0.4621   6.3833  11.8959

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.418e+00  1.007e+01   0.737  0.47262
gdp_per_capita      2.261e-03  7.593e-04   2.978  0.00939 **
I(gdp_per_capita^2) -1.956e-08  1.347e-08  -1.451  0.16729
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.347 on 15 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared:  0.7363,  Adjusted R-squared:  0.7012
F-statistic: 20.94 on 2 and 15 DF,  p-value: 4.55e-05
```

Example



The hat matrix

The hat matrix

If $\mathbf{x} = (1, x_0, x_1, \ldots, x_k)^T$, then the corresponding fitted model will be

$$y = \mathbf{x}^T \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

and the vector $\hat{\mathbf{y}} = (\hat{y}_1,\, \hat{y}_2, \ldots,\, \hat{y}_n)$ will be given by

$$\hat{\mathbf{y}} = X\hat{\beta} = \underbrace{X(X^T X)^{-1} X^T}_{\mathbf{H}} \mathbf{y}.$$

The matrix $\mathbf{H} := X(X^T X)^{-1} X^T$ is called the *hat matrix*.

Leverage

Since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, we can think of $\mathbf{H}_{ij}$ as expressing how much *leverage* the variable y_j exerts on the fitted value $\hat{y}_i$.

Residuals

Let $\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}}$ be the vector of residuals. Then $\mathbf{e} = (I - H)\mathbf{y}$, and hence $\mathrm{Cov}(\mathbf{e}) = \sigma^2(I - H)$.

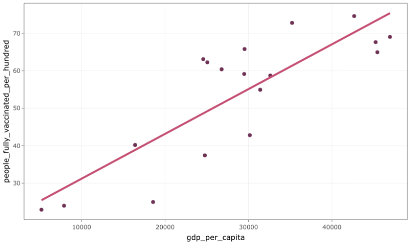
The hat matrix

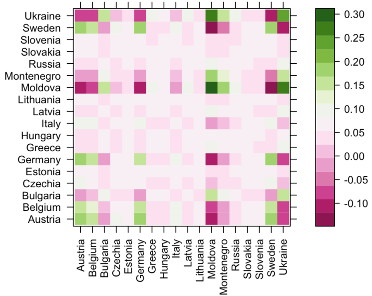
Leverage

Since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, we can think of $\mathbf{H}_{ij}$ as expressing how much *leverage* the variable y_j exerts on the fitted value $\hat{y}_i$.

Residuals

Let $\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}}$ be the vector of residuals. Then $\mathbf{e} = (I - H)\mathbf{y}$, and hence $\mathrm{Cov}(\mathbf{e}) = \sigma^2(I - H)$.





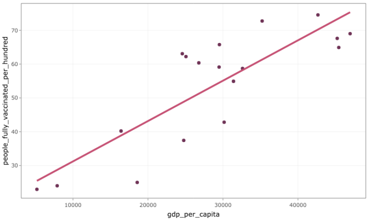
The hat matrix

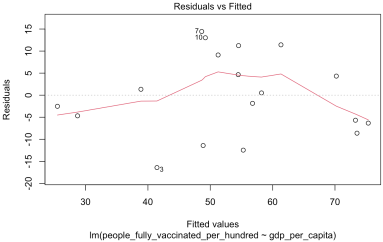
Model testing

If the model assumption is correct, we have

$$\mathrm{Cov}(\mathbf{e}, \hat{\mathbf{y}}) = \mathrm{Cov}\big((I - H)\mathbf{y}, H\mathbf{y}\big) = 0.$$

Consequently, a scatterplot of $\mathbf{e}$ vs $\hat{\mathbf{y}}$ should have no apparent slope or other pattern.





Properties of $\hat{\boldsymbol{\beta}}$

The expected value

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(X^TX)^{-1}X^T\mathbf{y}] = (X^TX)^{-1}X^T\mathbb{E}[\mathbf{y}] = (X^TX)^{-1}X^T\mathbb{E}[X\boldsymbol{\beta}+\boldsymbol{\varepsilon}] \\ &= (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}.\end{aligned}$$

In other words, if the model is correct, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

The covariance

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}}-\mathbb{E}[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\beta}}-\mathbb{E}[\hat{\boldsymbol{\beta}}])] = \mathrm{Var}(\hat{\boldsymbol{\beta}}-\mathbb{E}[\hat{\boldsymbol{\beta}}]) \\ &= \mathrm{Var}((X^TX)^{-1}X^T\mathbf{y}) = (X^TX)^{-1}X^T\,\mathrm{Var}(\mathbf{y})((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\sigma^2I((X^TX)^{-1}X^T)^T = \sigma^2(X^TX)^{-1}X^T((X^TX)^{-1}X^T)^T \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}.\end{aligned}$$

Gauss-Markov theorem

Theorem

$\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$.

Estimation of σ^2

Residuals

We let $\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}}$ be the vector of residuals.

The residual sum of squares

We have

$$\begin{aligned} SS_{Res} &= \sum (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})^T \overbrace{(\mathbf{y} - \hat{\mathbf{y}})}^{\mathbf{e}} = (\mathbf{y} - X\hat{\boldsymbol{\beta}})^T (\mathbf{y} - X\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^T\mathbf{y} - (X\hat{\boldsymbol{\beta}})^T\mathbf{y} - \mathbf{y}^T(X\hat{\boldsymbol{\beta}}) + (X\hat{\boldsymbol{\beta}})^T X\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T X^T\mathbf{y} - \mathbf{y}^T X\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T \underbrace{X^T X\hat{\boldsymbol{\beta}}}_{=X^T\mathbf{y}} = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T X\hat{\boldsymbol{\beta}} \end{aligned}$$

SS_{Res} has $n - (k+1)$ degrees of freedom. Also, one can show that $\mathbb{E}[SS_{Res}] = \sigma^2(n - (k+1))$.

The residual mean squared

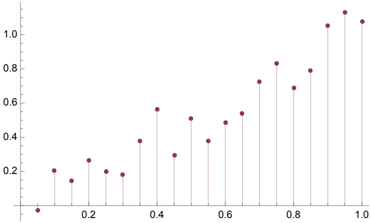
$$\hat{\sigma}^2 = MS_{Res} = \frac{SS_{Res}}{n - (k+1)}.$$

Note that $\hat{\sigma}^2$ is an unbiased estimate of σ^2. As in the simple linear regression case, this estimate depends on the model.

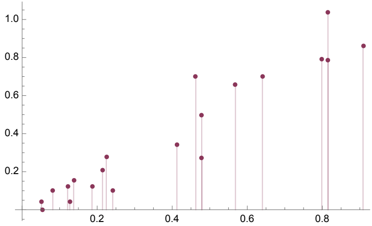
Random regression variables

Recall that the regression variables $x_1, x_2, \ldots$ can be either

1. non-random (e.g. data from planned experiments), or
2. random (common when we use already collected data).



non-random X



random X

So far, we have only covered case 1. In this case, we found the LS estimates by minimizing $S(\boldsymbol{\beta}) = \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$. Note that this means that we are minimizing the errors *at the sampled points*, and not for *general points*.
If the regression variables are random, we instead want to minimize

$$S(\boldsymbol{\beta}) = \mathbb{E}[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2] = \mathbb{E}[(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})].$$

Random regression variables

The "best" estimates

Let

$$S(\boldsymbol{\beta}) = \mathbb{E}[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2] = \mathbb{E}[(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})].$$

Differentiating, we obtain

$$\frac{dS(\boldsymbol{\beta})}{d\boldsymbol{\beta}} = -2\mathbb{E}[\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}] = -2\mathbb{E}[\mathbf{X}^T\mathbf{y}] + 2\mathbb{E}[\mathbf{X}^T\mathbf{X}]\boldsymbol{\beta}$$

and hence the "best" estimates are given by

$$\tilde{\boldsymbol{\beta}} := \mathbb{E}[\mathbf{X}^T\mathbf{X}]^{-1}\mathbb{E}[\mathbf{X}^T\mathbf{y}].$$

Equivalently,

$$(\tilde{\beta}_1, \ldots, \hat{\beta}_k) := \Sigma_{XX}^{-1}\Sigma_{XY} \quad \text{and} \quad \tilde{\beta}_0 := \mu_{\mathbf{y}} - \mu_{\mathbf{X}}(\hat{\beta}_1, \ldots, \hat{\beta}_k)^T.$$

An approximation of the "best" estimates

Note that since the joint distribution of $(\mathbf{X}, \mathbf{y})$ is in general not known, we cannot really use the above "best" estimeates.

For this reason, we let $X_* := (x_{ij})$ (so that $X = (1, X_*)$), and approximate these estimators by

$$\begin{cases} (\hat{\beta}_1, \ldots, \hat{\beta}_k) := \left((X_* - \bar{X}_*)^T(X_* - \bar{X}_*)\right)^{-1}(X_* - \bar{X}_*)^T(y - \bar{y}) \\ \hat{\beta}_0 := \bar{y} - \bar{X}(\hat{\beta}_1, \ldots, \hat{\beta}_k)^T \end{cases}$$

Commentes

→ The above estimators are identical to the LSE given for the non-random setting.

→ All previous results are applicable if $y \mid x \sim N(\beta_0 + \beta_1 x, \sigma^2)$ and the sampled x are independent (with some unknown distribution).