STAT 513 fa 2020 Lec 08 slides

Multiple linear regression

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Regression model

For data pairs $(Y_1, \mathbf{x}_1), \dots, (Y_n, \mathbf{x}_n)$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$, suppose

$$Y_i = f(x_{i1}, \dots, x_{ip}) + \varepsilon_i$$

for $i = 1, \dots, n$, where

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ are fixed vectors in $\mathbb{R}^p$
- $Y_1, \dots, Y_n$ are independent random variables
- $f : \mathbb{R}^p \to \mathbb{R}$ is an unknown function
- $\varepsilon_1, \dots, \varepsilon_n$ are iid errors with
  - $\mathbb{E}\varepsilon_i = 0$
  - $\operatorname{Var} \varepsilon_i = \sigma^2$

  for $i = 1, \dots, n$.

Goal: Estimate the unknown function f and the error variance σ^2.

Multiple linear regression model

For data pairs $(Y_1, \mathbf{x}_1), \dots, (Y_n, \mathbf{x}_n)$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$, suppose

$$Y_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{ip}\beta_p + \varepsilon_i$$

for $i = 1, \dots, n$, where

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ are fixed vectors in $\mathbb{R}^p$
- $Y_1, \dots, Y_n$ are independent random variables
- $\beta_0, \beta_1, \dots, \beta_p$ are unknown constants
- $\varepsilon_1, \dots, \varepsilon_n$ are iid errors with
  - $\mathbb{E}\varepsilon_i = 0$
  - $\operatorname{Var}\varepsilon_i = \sigma^2$

  for $i = 1, \dots, n$.

Goal: Estimate the unknown constants $\beta_0, \beta_1, \dots, \beta_p$ and the error variance σ^2.

Least-squares estimators of multiple linear regression coefficients

We define the least-squares estimators of $\beta_0, \beta_1, \dots, \beta_p$ as

$$(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p) = \underset{(\beta_0, \beta_1, \dots, \beta_p) \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]^2.$$

Expressions for $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are very complicated when $p > 1$. So use matrices!

Exercise: Define $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ so that the n equations

$$Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad i = 1, \dots, n$$

can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Least-squares estimators of multiple linear regression coefficients

Provided $\mathbf{X}^T\mathbf{X}$ is invertible, the function

$$Q_n(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

is (uniquely) minimized at

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

In the above, $\|\mathbf{x}\|_2^2 = x_1^2 + \cdots + x_d^2$ for $\mathbf{x} \in \mathbb{R}^d$ (squared Euclidean norm).

Exercise: Derive the above result using

$$\frac{\partial \mathbf{a}^T\mathbf{u}}{\partial \mathbf{u}} = \mathbf{a} \quad \text{and} \quad \frac{\partial \mathbf{u}^T\mathbf{A}\mathbf{u}}{\partial \mathbf{u}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{u}.$$

- The *fitted values* are the entries of the vector

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}.$$

- The *residuals* are the entries of the vector

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

Exercise:

1. Give the matrix $\mathbf{X}^T\mathbf{X}$ and the vector $\mathbf{X}^T\mathbf{Y}$ in the $p = 1$ case.
2. Verify on a toy dataset that when $p = 1$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ gives

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1\bar{x}_n \quad \text{and} \quad \hat{\beta}_1 = r_{xY}(s_Y/s_x).$$

Mean and covariance matrix of a random vector

Let $\mathbf{U} = (U_1, \dots, U_d)^T$ be a rvec and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)^T$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_{ij})_{1 \leq i,j \leq d}$ be the vector and matrix having entries such that

$$\begin{aligned} \mathbb{E}U_i &= \mu_i \quad \text{for } i = 1, \dots, d \\ \mathrm{Cov}(U_i, U_j) &= \boldsymbol{\Sigma}_{ij} \quad \text{for } 1 \leq i,j \leq d. \end{aligned}$$

Then $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the *mean vector* and the *covariance matrix* of $\mathbf{U}$.

- Use notation $\mathrm{Cov}(\mathbf{U}) = \boldsymbol{\Sigma}$.
- We have

$$\mathrm{Cov}(\mathbf{U}) = \mathbb{E}[(\mathbf{U} - \mathbb{E}\mathbf{U})(\mathbf{U} - \mathbb{E}\mathbf{U})^T].$$

Moments of linearly transformed random vector

Let $\mathbf{U} = (U_1, \dots, U_d)^T$ be a rvec and $\mathbf{a} = (a_1, \dots, a_d)^T$ a vector of reals. Then

$$\mathrm{Var}(\mathbf{a}^T\mathbf{U}) = \mathbf{a}^T\,\mathrm{Cov}(\mathbf{U})\mathbf{a}.$$

Moreover, if $\mathbf{A} = (A_{ij})_{1 \le i,j \le d}$ is a matrix of real numbers, then

$$\mathbb{E}(\mathbf{a} + \mathbf{A}\mathbf{U}) = \mathbf{a} + \mathbf{A}\mathbb{E}\mathbf{U}$$
$$\mathrm{Cov}(\mathbf{a} + \mathbf{A}\mathbf{U}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{U})\mathbf{A}^T.$$

Exercises:

- Derive the above.
- Find $\mathbb{E}\hat{\beta}$ and $\mathrm{Cov}(\hat{\beta})$.
- Find $\mathbb{E}\tilde{\mathbf{x}}_{\text{new}}^T\hat{\beta}$ and $\mathrm{Var}(\tilde{\mathbf{x}}_{\text{new}}^T\hat{\beta})$, where $\tilde{\mathbf{x}}_{\text{new}} = (1, \mathbf{x}_{\text{new}}^T)^T$.

Multivariate Normal distribution

The pdf of a rvec $\mathbf{U}$ having the *multivariate Normal distribution* with mean vector $\boldsymbol{\mu}$ and (invertible) covariance matrix $\boldsymbol{\Sigma}$ is given by

$$f(\mathbf{u};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})\right]$$

for all $\mathbf{u} \in \mathbb{R}^d$, where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

The mgf of $\mathbf{U}$ is given by

$$M_{\mathbf{U}}(\mathbf{t}) = \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right)$$

We write $\mathbf{U} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Exercise: Show that $\varepsilon \sim \text{Normal}(\mathbf{0},\sigma^2\mathbf{I}_n)$.

Distribution of linearly transformed multivariate Normal rvec

Let $\mathbf{U} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a $d \times 1$ random vector and let

$$\mathbf{V} = \mathbf{a} + \mathbf{A}\mathbf{U}$$

for some $r \times 1$ vector $\mathbf{a}$ and $r \times d$ matrix $\mathbf{A}$. Then

$$\mathbf{V} \sim \text{Normal}(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T).$$

Exercise: Derive the above using multivariate mgfs.

Sampling distribution results under Normal errors

If $\varepsilon_1, \dots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$, then

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim \text{Normal}(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \\ \mathbf{a}^T\hat{\boldsymbol{\beta}} &\sim \text{Normal}(\mathbf{a}^T\boldsymbol{\beta}, \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a} \cdot \sigma^2) \end{aligned}$$

for any vector $\mathbf{a} \in \mathbb{R}^{p+1}$, and

$$(n-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p-1}.$$

Moreover

$$\frac{\mathbf{a}^T\hat{\boldsymbol{\beta}} - \mathbf{a}^T\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \sim t_{n-p-1}.$$

An unbiased estimator of the variance is

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2.$$

Confidence intervals

For any $\mathbf{a} \in \mathbb{R}^{p+1}$, a $(1-\alpha)100\%$ confidence interval for $\mathbf{a}^T\beta$ is given by

$$\mathbf{a}^T\hat{\beta} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}.$$

We can choose $\mathbf{a}$ to build CIs of interest.

Exercise: Show that a $(1-\alpha) \times 100\%$ CI for β_j is given by

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\hat{\Omega}_{jj}^{1/2}, \quad j = 1, \dots, p,$$

where $\hat{\Omega}_{jj}$ is the (j,j) element of $(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}$.

Tests of hypotheses

For some $\mathbf{a} \in \mathbb{R}^{p+1}$, consider tests comparing $\mathbf{a}^T\beta$ to a null value a^* and define

$$T_n = \frac{\mathbf{a}^T\hat{\beta} - a^*}{\hat{\sigma}\sqrt{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}}.$$

We have the following:

H_0	H_1	Reject H_0 at α iff	p-value
$\mathbf{a}^T\beta \leq a^*$	$\mathbf{a}^T\beta > a^*$	$T_n > t_{n-p-1,\alpha}$	$1 - F_{t_{n-p-1}}(T_n)$
$\mathbf{a}^T\beta \geq a^*$	$\mathbf{a}^T\beta < a^*$	$T_n < -t_{n-p-1,\alpha}$	$F_{t_{n-p-1}}(T_n)$
$\mathbf{a}^T\beta = a^*$	$\mathbf{a}^T\beta \neq a^*$	$\lvert T_n\rvert > t_{n-p-1,\alpha/2}$	$2[1 - F_{t_{n-p-1}}(\lvert T_n\rvert)]$

Prediction interval for a new observation

A $(1-\alpha)\times 100\%$ prediction interval for Y_{new} of a new obs. $(Y_{\text{new}}, \mathbf{x}_{\text{new}})$ is given by

$$\tilde{\mathbf{x}}_{\text{new}}^T \hat{\boldsymbol{\beta}} \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{1 + \tilde{\mathbf{x}}_{\text{new}}^T(\mathbf{X}^T\mathbf{X})^{-1}\tilde{\mathbf{x}}_{\text{new}}},$$

where $\tilde{\mathbf{x}}_{\text{new}} = (1, \mathbf{x}_{\text{new}}^T)^T$.

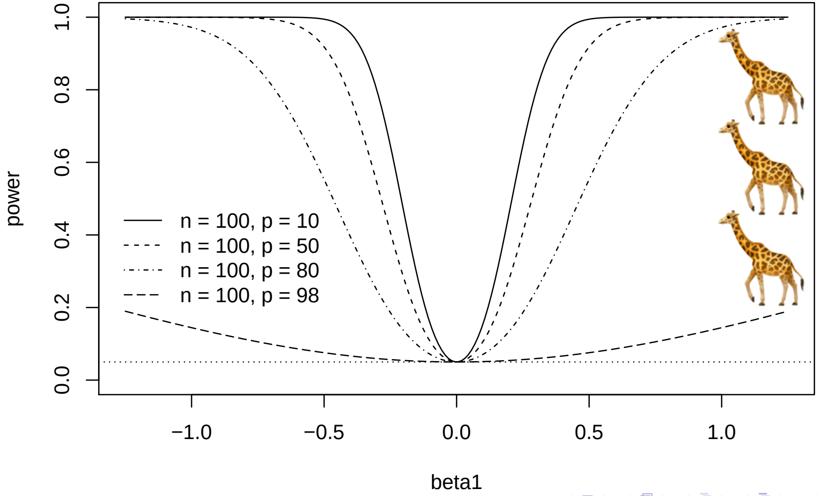
Exercise: Derive the above using the distribution of $\hat{\varepsilon}_{\text{new}} = Y_{\text{new}} - \tilde{\mathbf{x}}_{\text{new}}^T\hat{\boldsymbol{\beta}}$.

Exercise: Run `data(trees)` in R and fit the model

$$\text{Volume}_i = \beta_0 + \beta_1 \cdot \text{Girth}_i + \beta_2 \cdot \text{Height}_i + \varepsilon_i, \quad i = 1, \dots, n.$$

1. Build a 99% CI for β_1, the coefficient for girth.
2. Build a 99% CI for β_2, the coefficient for height.
3. Get the p-value for testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ and interpret it.
4. Get the p-value for testing H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$ and interpret it.
5. Build a 95% CI for the average volume of trees with girth 15 and height 70.
6. Build a 95% PI for the volume of a tree with girth 15 and height 70.

Some power curves of the test for H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.



Likelihood of multiple linear regression under Normal errors

Let $\varepsilon_1, \dots, \varepsilon_n \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$. Then

- The likelihood function of $\boldsymbol{\beta}$ and σ^2 is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2\right].$$

- The log-likelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Exercise: For any $r \in \{1, \dots, p\}$, show that the likelihood ratio test of

$$H_0\colon \beta_{r+1} = \cdots = \beta_p = 0 \text{ vs } H_1\colon \beta_j \neq 0 \text{ for some } j \in \{r+1, \dots, p\}$$

is of the form

$$\text{Reject } H_0 \text{ iff } \left[\frac{\|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\beta}_{\mathcal{R}}\|_2^2}{\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2}\right]^{-n/2} < c,$$

where

- $\mathbf{X}_{\mathcal{R}}$ is the matrix formed by the first $r+1$ columns of $\mathbf{X}$.
- $\hat{\beta}_{\mathcal{R}} = (\mathbf{X}_{\mathcal{R}}^T\mathbf{X}_{\mathcal{R}})^{-1}\mathbf{X}_{\mathcal{R}}^T\mathbf{Y}$.

Full-reduced model F-test

The *full-reduced model* F-test of

$$H_0\colon \beta_{r+1} = \cdots = \beta_p = 0 \text{ vs } H_1\colon \beta_j \neq 0 \text{ for some } j \in \{r+1, \dots, p\}$$

for any $r \in \{1, \dots, p\}$ is

$$\text{Reject } H_0 \text{ iff } \frac{(\mathrm{SSE}_{\mathrm{Red}} - \mathrm{SSE}_{\mathrm{Full}})/(p-r)}{\mathrm{SSE}_{\mathrm{Full}}/(n-p-1)} > F_{p-r,n-p-1,\alpha},$$

where $\mathrm{SSE}_{\mathrm{Red}} = \|\mathbf{Y} - \mathbf{X}_{\mathcal{R}}\hat{\beta}_{\mathcal{R}}\|_2^2$ and $\mathrm{SSE}_{\mathrm{Full}} = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_2^2$.

Exercise: Use the result that under H_0

$$\begin{aligned} \mathrm{SSE}_{\mathrm{Full}}/\sigma^2 &\sim \chi^2_{n-p-1} \\ (\mathrm{SSE}_{\mathrm{Red}} - \mathrm{SSE}_{\mathrm{Full}})/\sigma^2 &\sim \chi^2_{p-r} \end{aligned}$$

and the independence of these quantities to show that this is the size-α LRT.

Exercise: Run `data(swiss)` in R and consider the model

$$\text{Fert}_i = \beta_0 + \beta_{\text{Ag}}\,\text{Ag}_i + \beta_{\text{Ex}}\,\text{Ex}_i + \beta_{\text{Ed}}\,\text{Ed}_i + \beta_{\text{Cath}}\,\text{Cath}_i + \beta_{\text{InfM}}\,\text{InfM}_i + \varepsilon_i$$

for $i = 1, \dots, 47$.

Do the following:

1. Get LS estimators in the full model.
2. Compute SSE_{Full}.
3. Fit reduced model after omitting `Examination` and `Education`.
4. Compute SSE_{Red}.
5. Compute full-reduced model F statistic H_0: $\beta_{\text{Examination}} = \beta_{\text{Education}} = 0$.
6. Compute the p-value for the full-reduced model F-test.

Overall test of significance for the linear regression model

The size-α *overall F-test of significance* is the test of the hypotheses

$$H_0\colon \beta_1 = \cdots = \beta_p = 0 \text{ versus } H_1\colon \beta_j \neq 0 \text{ for some } j \in \{1, \ldots, p\}.$$

which has rejection rule

$$\text{Reject } H_0 \text{ iff } \frac{\text{SSM}/p}{\text{SSE}/(n-p-1)} > F_{p,n-p-1,\alpha},$$

where $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ and $\text{SSM} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$.

Exercise: Show that this is just the full-reduced model F-test with $r = 0$.

Can be reformulated as

$$\text{Reject } H_0 \text{ iff } \frac{\text{MSM}}{\text{MSE}} > F_{p,n-p-1,\alpha}.$$

Analysis of Variance (ANOVA) table

	df	SS	MS	F_n	p-value
Model	p	SSM	MSM	F_n	$1 - F_{F_{p,n-p-1}}(F_n)$
Error	$n - p - 1$	SSE	MSE		
Total	$n - 1$	SST			

An oft-used tabulation of the quantities involved in the overall F-test.

Exercise: Fill out the ANOVA table for our model for the `swiss` data.

The *coefficient of determination* is defined as the quantity

$$R^2 = \frac{\text{SSM}}{\text{SST}}.$$