

13 Lecture 13: Feb 23

Last time

- Lab 2 review
- Multiple linear regression

Today

- HW1 review next week
- HW2 posted, due March 4th
- Inference of MLR
- more review on probability

Matrix formulation of MLR

Let a vector for p observed independent variables for individual i be defined by

$$\mathbf{x}_{i\cdot} = (1, x_{i1}, x_{i2}, \dots, x_{ip}).$$

The MLR model for Y_1, \dots, Y_n is given by

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_p X_{1p} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_p X_{2p} + \epsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_p X_{np} + \epsilon_n \end{aligned}$$

This system of n equations can be expressed using matrices:

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

where

- \mathbf{Y} denotes a response vector of size $n \times 1$
- \mathbf{X} denotes a design matrix of size $n \times (p + 1)$
- $\boldsymbol{\beta}$ denotes a vector of regression parameters of size $(p + 1) \times 1$
- $\boldsymbol{\epsilon}$ denotes an error vector of size $n \times 1$

Here, the error vector $\boldsymbol{\epsilon}$ is assumed to follow a multivariate normal distribution with variance-covariance matrix $\sigma^2 \mathbf{I}_n$. For individual i ,

$$y_i = \mathbf{x}_{i\cdot} \boldsymbol{\beta} + \epsilon_i.$$

Some simplified expressions: (\mathbf{a} is a known $p \times 1$ vector)

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ \text{Var}(\hat{\boldsymbol{\beta}}) &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \boldsymbol{\Sigma} \\ \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) &= MS[E] (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \hat{\boldsymbol{\Sigma}} \\ \widehat{\text{Var}}(\mathbf{a}^T \hat{\boldsymbol{\beta}}) &= \mathbf{a}^T \hat{\boldsymbol{\Sigma}} \mathbf{a}\end{aligned}$$

Question: what are the dimensions of each of these quantities?

- $(\mathbf{X}^T \mathbf{X})^{-1}$ may be verbalized as “x transposed x inverse”
- $\hat{\boldsymbol{\Sigma}}$ is the estimated variance-covariance matrix for the estimate of the regression parameter vector $\hat{\boldsymbol{\beta}}$
- \mathbf{X} is assumed to be of full *rank*.

Some more simplified expressions:

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \\ &= \mathbf{H} \mathbf{Y} \\ \hat{\boldsymbol{\epsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= (\mathbf{I} - \mathbf{H}) \mathbf{Y}\end{aligned}$$

- $\hat{\mathbf{Y}}$ is called the vector of fitted or predicted values
- $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is called the hat matrix
- $\hat{\boldsymbol{\epsilon}}$ is the vector of residuals

For the Duncan’s data example on income, education and prestige, with $p = 2$ independent variables and $n = 45$ observations,

$$\mathbf{X} = \begin{bmatrix} 1 & 62 & 86 \\ 1 & 72 & 76 \\ \vdots & \vdots & \vdots \\ 1 & 8 & 32 \end{bmatrix}$$

and

$$\begin{aligned}
\mathbf{X}^T \mathbf{X} &= \begin{bmatrix} 45 & 1884 & 2365 \\ 1884 & 105148 & 122197 \\ 2365 & 122197 & 163265 \end{bmatrix} \\
(\mathbf{X}^T \mathbf{X})^{-1} &= \begin{bmatrix} 0.10211 & -0.00085 & -0.00084 \\ -0.00085 & 0.00008 & -0.00005 \\ -0.00084 & -0.00005 & 0.00005 \end{bmatrix} \\
(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} &= \begin{bmatrix} -6.0646629 \\ 0.5987328 \\ 0.5458339 \end{bmatrix} = ? \\
SS[E] &= \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}) = 7506.7 \\
MS[E] &= \frac{SS[E]}{df} = \frac{7506.7}{45 - 2 - 1} = 178.73 \\
\hat{\boldsymbol{\Sigma}} &= MS[E] (\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} 18.249481 & -0.151845008 & -0.150706025 \\ -0.151845 & 0.014320275 & -0.008518551 \\ -0.150706 & -0.008518551 & 0.009653582 \end{bmatrix}
\end{aligned}$$

Multiple correlation, JF 5.2.3

The sums of squares in multiple regression are defined in the same manner as in SLR:

$$\begin{aligned}
TSS &= \sum (Y_i - \bar{Y})^2 \\
RegSS &= \sum (\hat{Y}_i - \bar{Y})^2 \\
RSS &= \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{\epsilon}_i^2
\end{aligned}$$

Not surprisingly, we have a similar analysis of variance for the regression:

$$TSS = RegSS + RSS$$

The squared multiple correlation R^2 , representing the proportion of variation in the response variable captured by the regression, is defined in terms of the sums of squares:

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Because there are several slope coefficients, potentially with different signs, the *multiple correlation coefficient* is, by convention, the positive square root of R^2 . The multiple correlation is also interpretable as the simple correlation between the fitted and observed Y values, i.e. $r_{\hat{Y}Y}$.

Adjusted- R^2

Because the multiple correlation can only rise, never decline, when explanatory variables are added to the regression equation (HW1), investigators sometimes penalize the value of R^2 by a “correction” for degrees of freedom. The corrected (or “adjusted”) R^2 is defined as:

$$\begin{aligned} R_{adj}^2 &= 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}} \\ &= 1 - \left[\frac{(1 - R^2)(n - 1)}{n - p - 1} \right] \end{aligned}$$

Confidence intervals

Confidence intervals and hypothesis tests for individual coefficients closely follow the pattern of simple-regression analysis:

1. substitute an estimate of the error variance (MSE) for the unknown σ^2 into the variance term of $\hat{\beta}_i$
2. find the estimated standard error of a slope coefficient $\widehat{SE}(\hat{\beta}_i)$
3. $t = \frac{\hat{\beta}_i - \beta_i}{\widehat{SE}(\hat{\beta}_i)}$ follows a t -distribution with degrees of freedom as associated with SSE.

Therefore, we can construct the $100(1 - \alpha)\%$ confidence interval for a single slope parameter by (why?):

$$\hat{\beta}_i \pm t(n - p - 1, \alpha/2) \widehat{SE}(\hat{\beta}_i)$$

Hand-waving proof:

Hypothesis tests

We first test the null hypothesis that all population regression slopes are 0:

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

The test statistics,

$$F = \frac{RegSS/p}{RSS/(n - p - 1)}$$

follows an F -distribution with p and $n - p - 1$ degrees of freedom.

We can also test a null hypothesis about a *subset* of the regression slopes, e.g.,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0.$$

Or more generally, test the null hypothesis

$$H_0 : \beta_{q_1} = \beta_{q_2} = \cdots = \beta_{q_k} = 0$$

where $0 \leq q_1 < q_2 < \cdots < q_k \leq p$ is a subset of k indices. To get the F -statistic for this case, we generally perform the following steps:

1. Fit the *full* (“unconstrained”) model, in other words, model that provides context for H_0 . Record SSR_{full} and the associated df_{full}
2. Fit the *reduced* (“constrained”) model, in other words, full model constrained by H_0 . Record SSR_{red} and the associated df_{red}
3. Calculate the F-statistic by

$$F = \frac{[SSR_{red} - SSR_{full}]/(df_{red} - df_{full})}{SSR_{full}/df_{full}}$$

4. Find p -value (the probability of observing an F-statistic that is at least as high as the value that we obtained) by consulting an F-distribution with numerator $df(ndf) = df_{red} - df_{full}$ and denominator $df(ddf) = df_{full}$. Notation: $F_{ndf,ddf}$, see Figure 13.1.

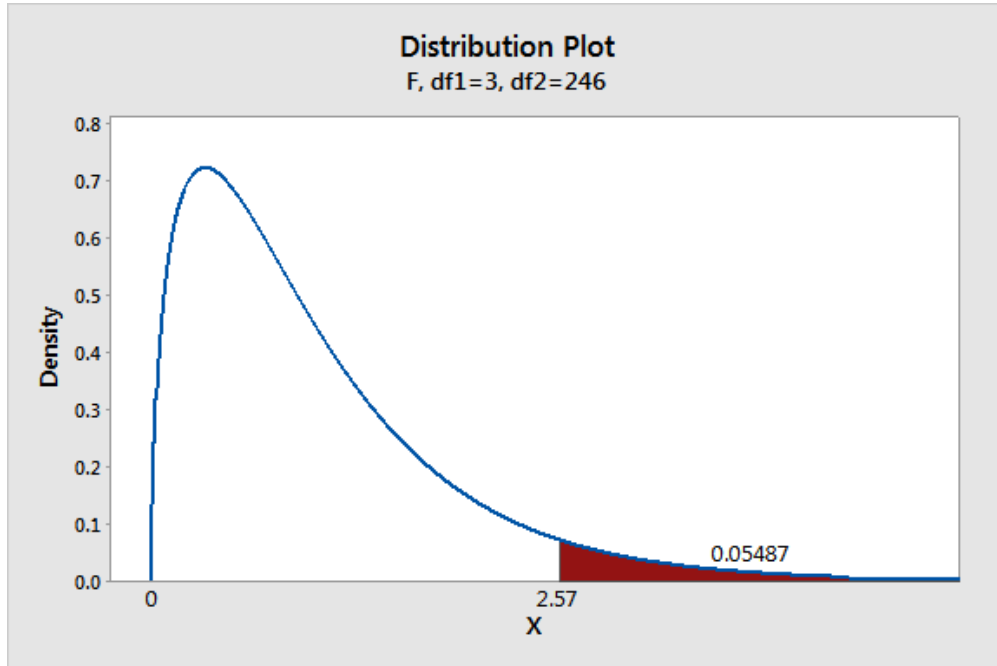


Figure 13.1: An example for p -value for F-statistic value 2.57 with an $F_{3,246}$ distribution

A little more background review

Reference:

- Statistical Inference, 2nd Edition, by George Casella & Roger L. Berger
- [Review of Probability Theory](#) by Arian Maleki and Tom Do

Chi-square, t-, and F-Distributions

Let $Z_1, Z_2, \dots, Z_k \stackrel{iid}{\sim} N(0, 1)$, then $X^2 \equiv Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi_k^2$ (with k degrees of freedom).
If $X \sim \chi_k^2$

$$\begin{aligned}\mathbf{E}(X) &= k \\ \mathbf{Var}(X) &= 2k.\end{aligned}$$

Student's t versus χ^2

If $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

When σ is unknown,

$$\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}, \quad \text{where } \hat{\sigma} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}}.$$

Note that

$$\begin{aligned}\frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{1}{\frac{\hat{\sigma}}{\sigma}} \\ &= Z \cdot \frac{1}{\sqrt{\frac{\sum (X_i - \bar{X})^2}{(n-1)\sigma^2}}} \\ &= \frac{Z}{\sqrt{\frac{\chi_{n-1}^2}{n-1}}}\end{aligned}$$

F versus χ^2

$$F_{ndf,ddf} \equiv \frac{\chi_{ndf}^2/ndf}{\chi_{ddf}^2/ddf}$$

t versus F

$$\begin{aligned}
t_k &= \frac{Z}{\sqrt{\chi_k^2/k}} \\
&= \frac{\sqrt{\chi_1^2/1}}{\sqrt{\chi_k^2/k}} \\
&= \sqrt{F_{1,k}}
\end{aligned}$$

or, in other words, $t_k^2 = F_{1,k}$

Random vectors and matrices

The cdf for random vector

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \text{ is } F_{\mathbf{Y}}(\mathbf{y}) = \Pr(Y_1 \leq y_1, Y_2 \leq y_2, \dots, Y_n \leq y_n)$$

If a joint pdf exists, then $f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{Y}}(y_1, \dots, y_n)$ and

$$F_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \dots \int_{-\infty}^{y_n} f_{\mathbf{Y}}(\mathbf{t}) d\mathbf{t}$$

Moments

$$\begin{aligned}
\mathbf{E}(\mathbf{Y}) &= \boldsymbol{\mu}_{\mathbf{Y}} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \\
\mathbf{Var}(\mathbf{Y}) &= \mathbf{E}((\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^T) \\
&= \mathbf{E} \left(\begin{bmatrix} (Y_1 - \mu_1)^2 & (Y_1 - \mu_1)(Y_2 - \mu_2) & \dots \\ (Y_2 - \mu_2)(Y_1 - \mu_1) & (Y_2 - \mu_2)^2 & \dots \\ \dots & & \end{bmatrix} \right) \\
&= \mathbf{E}([(Y_i - \mu_i)(Y_j - \mu_j), i = 1, 2, \dots, n, j = 1, 2, \dots, n]) \\
&= (\sigma_{ij})_{i=1,2,\dots,n; j=1,2,\dots,n}
\end{aligned}$$

where $\sigma_{ij} = Cov(Y_i, Y_j)$

Linear functions

Let $\mathbf{X} \in \mathbb{R}^{k \times 1}$, $\mathbf{Y} \in \mathbb{R}^{n \times 1}$ and $\mathbf{A} \in \mathbb{R}^{k \times 1}$, $\mathbf{B} \in \mathbb{R}^{k \times n}$ be non-random, then

$$\begin{aligned}
\mathbf{X} &= \mathbf{A} + \mathbf{B} \mathbf{Y} \\
&\quad \begin{matrix} k \times 1 & k \times 1 & k \times n & n \times 1 \end{matrix} \\
\mathbf{E}(\mathbf{X}) &= \mathbf{A} + \mathbf{B} \mathbf{E}(\mathbf{Y}) \\
\mathbf{Var}(\mathbf{X}) &= \mathbf{B} \mathbf{Var}(\mathbf{Y}) \mathbf{B}^T
\end{aligned}$$

Sums of random vectors

$$\begin{aligned}\mathbf{X}_{n \times 1} &= \mathbf{Y}_{n \times 1} + \mathbf{Z}_{n \times 1} \\ \mathbf{E}(\mathbf{X}) &= \mathbf{E}(\mathbf{Y}) + \mathbf{E}(\mathbf{Z}) = \mathbf{E}(\mathbf{Y} + \mathbf{Z})\end{aligned}$$

Note that there is no independence assumed above.

$$\mathbf{Var}(\mathbf{X}) = \mathbf{Var}(\mathbf{Y} + \mathbf{Z}) = \mathbf{Var}(\mathbf{Y}) + \mathbf{Var}(\mathbf{Z}) + Cov(\mathbf{Y}, \mathbf{Z}) + Cov(\mathbf{Z}, \mathbf{Y})$$

If \mathbf{Y}, \mathbf{Z} are uncorrelated, then $\mathbf{Var}(\mathbf{X}) = \mathbf{Var}(\mathbf{Y}) + \mathbf{Var}(\mathbf{Z})$