Chapter 5 - Linear Model Theory

Joshua French

To open this information in an interactive Colab notebook, click the Open in Colab graphic below.

Basic theoretical results for linear models

In this chapter we discuss many basic theoretical results for linear models.

We assume the responses can be modeled as

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,-1} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

or using matrix formulation, as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

Standard assumptions

We assume that the components of our linear model have the characteristics previously described in Chapter 3. We also need to make several specific assumptions about the errors.

Error Assumption 1

The mean of the errors is zero conditional on the value of the regressors.

This means that

$$E(\epsilon_i \mid \mathbb{X} = \mathbf{x}_i) = 0, i = 1, 2, \ldots, n,$$

or using matrix notation,

$$E(\epsilon \mid \mathbf{X}) = 0_{n\times 1}.$$

where "$\mid \mathbf{X}$" is notation meaning "conditional on knowing the regressor values for all observations".

Error Assumption 2

The errors have constant variances and are uncorrelated, conditional on knowing the regressors, i.e.,

$$\text{var}(\epsilon_i \mid \mathbb{X} = \mathbf{x}_i) = \sigma^2, \quad i = 1, 2, \dots, n.$$

and

$$\text{cov}(\epsilon_i, \epsilon_j \mid \mathbf{X}) = 0, \quad i, j = 1, 2, \dots, n, \quad i \neq j.$$

In matrix notation, this is stated as

$$\text{var}(\epsilon \mid \mathbf{X}) = \sigma^2 \mathbf{I}_{n\times n}.$$

Error Assumption 3

The errors are identically distributed. This may be written as

$$\epsilon_i \sim F, i = 1, 2, \dots, n,$$

where F is some arbitrary distribution.

Error Assumption 4

In practice, it is common to assume the errors have a normal (Gaussian) distribution.

Assumptions 1-4 combined

Two uncorrelated normal random variables are also independent (but this is not generally true for other distributions).

Putting assumptions 1-4 together, we have that

$$\epsilon_1, \epsilon_2, \dots, \epsilon_n \mid \mathbf{X} \overset{i.i.d.}{\sim} \mathsf{N}(0, \sigma^2),$$

or using matrix notation,

$$\epsilon \mid \mathbf{X} \sim \mathsf{N}(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_{n\times n}).$$

In summary, our error assumptions are:

1. $E(\epsilon_i \mid \mathbb{X} = \mathbf{x}_i) = 0$ for $i = 1, 2, \dots, n$.
2. $\mathrm{var}(\epsilon_i \mid \mathbb{X} = \mathbf{x}_i) = \sigma^2$ for $i = 1, 2, \dots, n$.
3. $\mathrm{cov}(\epsilon_i, \epsilon_j \mid \mathbf{X}) = 0$ for $i \neq j$ with $i, j = 1, 2, \dots, n$.
4. ϵ_i has a normal distribution for $i = 1, 2, \dots, n$.

Summary of results

Combining these results with our linear model, we have:

1. $\mathbf{y} \mid \mathbf{X} \sim \mathsf{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}_{n\times n})$.
2. $\hat{\beta} \mid \mathbf{X} \sim \mathsf{N}(\beta, \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1})$.
3. $\hat{\epsilon} \mid \mathbf{X} \sim \mathsf{N}(\mathbf{0}_{n\times 1}, \sigma^2 (\mathbf{I}_{n\times n} - \mathbf{H}))$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
4. $\hat{\beta}$ has the minimum variance among all unbiased estimators of β with the additional assumptions that the model is correct and $\mathbf{X}$ is full-rank.

We prove these results in the sections below. To simplify the derivations below, we let $\mathbf{I} = \mathbf{I}_{n\times n}$.

Results for y

For our given linear model and under the assumptions summarized previously, our response variable has mean

$$E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta.$$

Proof:

$$\begin{aligned}
E(\mathbf{y}|\mathbf{X}) &= E(\mathbf{X}\beta + \epsilon|\mathbf{X}) && \text{(by definition)} \\
&= E(\mathbf{X}\beta|\mathbf{X}) + E(\epsilon|\mathbf{X}) && \text{(linearity of expectation)} \\
&= E(\mathbf{X}\beta|\mathbf{X}) + \mathbf{0}_{n\times 1} && \text{(by assumption about } \boldsymbol{\epsilon}) \\
&= \mathbf{X}\beta && \text{(since } \mathbf{X} \text{ and } \beta \text{ are constant)}
\end{aligned}$$

For the variance of the response:

$$\mathrm{var}(\mathbf{y} \mid \mathbf{X}) = \sigma^2 \mathbf{I}.$$

Proof:

$$\begin{aligned}
\mathrm{var}(\mathbf{y}|\mathbf{X}) &= \mathrm{var}(\mathbf{X}\beta + \epsilon|\mathbf{X}) && \text{(by definition)} \\
&= \mathrm{var}(\epsilon|\mathbf{X}) && \text{($\mathbf{X}\beta$ is constant)} \\
&= \sigma^2 \mathbf{I}. && \text{(by assumption)}
\end{aligned}$$

The response variable has the following distribution:

$$\mathbf{y} \mid \mathbf{X} \sim \mathsf{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I}).$$

Proof:

We have shown that:

- $E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta$.
- $\mathrm{var}(\mathbf{y}\mathbf{X}) = \sigma^2 \mathbf{I}$.

Since $\mathbf{y}$ is a linear function of the multivariate normal vector ϵ, then $\mathbf{y}$ must also have a multivariate normal distribution.

Results for $\hat{\beta}$

The OLS estimator for β is

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

This is an unbiased estimator for β, i.e.,

$$E(\hat{\beta} \mid \mathbf{X}) = \beta.$$

Proof:

We previously derived the following results,

$$E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta.$$

$$\mathrm{var}(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Then,

$$\begin{aligned}
E(\hat{\beta}|\mathbf{X}) &= E((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}) && \text{(substitute OLS formula)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y}|\mathbf{X}) && \text{(factor non-random terms)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} && \text{(above result)} \\
&= \mathbf{I}_{p\times p}\beta && \text{(property of inverse matrices)} \\
&= \beta
\end{aligned}$$

The OLS estimator $\hat{\beta}$ has variance

$$\mathrm{var}(\hat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Proof:

$$\begin{aligned}
\mathrm{var}(\hat{\beta}|\mathbf{X}) &= \mathrm{var}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}|\mathbf{X}) && \text{(by OLS formula)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{var}(\mathbf{y}|\mathbf{X})((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T && \text{(pull constants out of variance)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{var}(\mathbf{y}|\mathbf{X})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && \text{(simplification)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && \text{(previous result)} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (\sigma^2 \text{ is a scalar}) \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} && \text{(simplification)}
\end{aligned}$$

The OLS estimator $\hat{\beta}$ has the following distribution:

$$\hat{\beta} \mid \mathbf{X} \sim \mathsf{N}(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

Proof:

Since $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a linear combination of $\mathbf{y}$, and $\mathbf{y}$ is a multivariate normal random vector, then $\hat{\beta}$ is also a multivariate normal random vector. Using the previous two results for the expectation and variance,

$$\widehat{\beta}|\mathbf{X} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

Results for the residuals

The residual vector can be expressed in various equivalent ways, such as

$$\begin{aligned}
\hat{\epsilon} &= \mathbf{y} - \hat{\mathbf{y}} \\
&= \mathbf{y} - \mathbf{X}\widehat{\beta}.
\end{aligned}$$

The **hat** matrix is denoted as:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Thus, using the substitution $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and the definition for $\mathbf{H}$, we see that:

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

The hat matrix is an important theoretical matrix, as it projects $\mathbf{y}$ into the space spanned by the vectors in $\mathbf{X}$.

The hat matrix $\mathbf{H}$ is symmetric and idempotent.

Proof:

Notice that:

$$\begin{aligned}
\mathbf{H}^T &= (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T && \text{(definition of } \mathbf{H}\text{)} \\
&= (\mathbf{X}^T)^T((\mathbf{X}^T\mathbf{X})^{-1})^T\mathbf{X}^T && \text{(apply transpose to matrix product)} \\
&= \mathbf{X}((\mathbf{X}^T\mathbf{X})^T)^{-1}\mathbf{X}^T && \text{(simplification, reversibility of inverse and transpose)} \\
&= \mathbf{X}(\mathbf{X}^T(\mathbf{X}^T)^T)^{-1}\mathbf{X}^T && \text{(apply transpose to matrix product)} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T && \text{(simplification)} \\
&= \mathbf{H}
\end{aligned}$$

Thus, $\mathbf{H}$ is symmetric.

Additionally:

$$
\begin{aligned}
\mathbf{H}\mathbf{H} &= (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) && \text{(definition of } \mathbf{H}\text{)} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^t && \text{(associative property of matrices)} \\
&= \mathbf{X}\mathbf{I}_{p\times p}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T && \text{(property of inverse matrices)} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T && \text{(simplification)} \\
&= \mathbf{H}
\end{aligned}
$$

Therefore, $\mathbf{H}$ is idempotent.

The matrix $\mathbf{I}-\mathbf{H}$ is symmetric and idempotent.

Proof:

First, notice that:

$$
\begin{aligned}
(\mathbf{I}-\mathbf{H})^T &= \mathbf{I}^T - \mathbf{H}^T && \text{(transpose to matrix sum)} \\
&= \mathbf{I}-\mathbf{H} && \text{(since } \mathbf{I} \text{ and } \mathbf{H} \text{ are symmetric)}
\end{aligned}
$$

Thus, $\mathbf{I}-\mathbf{H}$ is symmetric.

Next:

$$
\begin{aligned}
(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) &= \mathbf{I} - 2\mathbf{H} + \mathbf{H}\mathbf{H} && \text{(transpose to matrix sum)} \\
&= \mathbf{I} - 2\mathbf{H} + \mathbf{H} && \text{(since H is idempotent)} \\
&= \mathbf{I}-\mathbf{H} && \text{(simplification)}
\end{aligned}
$$

Thus, $\mathbf{I}-\mathbf{H}$ is idempotent.

Under the assumptions we discussed previously, the residuals have mean

$$
E(\hat{\epsilon} \mid \mathbf{X}) = \mathbf{0}_{n\times 1}.
$$

Proof:

$$
\begin{aligned}
E(\hat{\epsilon}|\mathbf{X}) &= E((\mathbf{I}-\mathbf{H})\mathbf{y}|\mathbf{X}) \\
&= (\mathbf{I}-\mathbf{H})E(\mathbf{y}|\mathbf{X}) && \text{(}\mathbf{I}-\mathbf{H}\text{ is non-random)} \\
&= (\mathbf{I}-\mathbf{H})\mathbf{X}\beta && \text{(earlier result)} \\
&= \mathbf{X}\beta - \mathbf{X}\beta && \text{(distribute the product)} \\
&= \mathbf{X}\beta - \mathbf{X}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta && \text{(definition of H)} \\
&= \mathbf{X}\beta - \mathbf{X}\mathbf{I}_{p\times p}\beta && \text{(property of inverse matrix)} \\
&= \mathbf{X}\beta - \mathbf{X}\beta && \text{(simplification)} \\
&= \mathbf{0}_{n\times 1} && \text{(simplification)}
\end{aligned}
$$

The residuals have variance

$$\mathrm{var}(\hat{\epsilon} \mid \mathbf{X}) = \sigma^2(\mathbf{I}-\mathbf{H})$$

Proof:

$$
\begin{aligned}
\mathrm{var}(\hat{\epsilon}|\mathbf{X}) &= \mathrm{var}((\mathbf{I}-\mathbf{H})\mathbf{y}|\mathbf{X}) \\
&= (\mathbf{I}-\mathbf{H})\mathrm{var}(\mathbf{y}|\mathbf{X})(\mathbf{I}-\mathbf{H})^T && \text{(}\mathbf{I}-\mathbf{H}\text{ is nonrandom)} \\
&= (\mathbf{I}-\mathbf{H})\sigma^2(\mathbf{I}-\mathbf{H})^T && \text{(earlier result)} \\
&= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) && \text{(}\mathbf{I}-\mathbf{H}\text{ is symmetric)} \\
&= \sigma^2(\mathbf{I}-\mathbf{H}) && \text{(}\mathbf{I}-\mathbf{H}\text{ is idempotent)}
\end{aligned}
$$

The residuals have the following distribution:

$$\hat{\epsilon} \mid \mathbf{X} \sim \mathsf{N}(\mathbf{0}_{n\times 1}, \sigma^2(\mathbf{I}-\mathbf{H})).$$

Proof:

Since $\hat{\epsilon}$ is a linear combination of multivariate normal vectors, and using previous results, it has mean $\mathbf{0}_{n\times 1}$ and variance matrix $\sigma^2(\mathbf{I}-\mathbf{H})$.

The RSS can be represented as

$$RSS = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}.$$

Proof:

$$
\begin{aligned}
RSS &= \hat{\epsilon}^T\hat{\epsilon} && \text{(matrix representation of RSS)}\\
&= ((\mathbf{I}-\mathbf{H})\mathbf{y})^T(\mathbf{I}-\mathbf{H})\mathbf{y} && \text{(previous result)}\\
&= \mathbf{y}^T(\mathbf{I}-\mathbf{H})^T(\mathbf{I}-\mathbf{H})\mathbf{y} && \text{(apply transpose)}\\
&= \mathbf{y}^T(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\mathbf{y} && (\mathbf{I}-\mathbf{H} \text{ is symmetric)}\\
&= \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y} && (\mathbf{I}-\mathbf{H} \text{ is idempotent)}
\end{aligned}
$$

The Gauss-Markov Theorem

Suppose we will fit the regression model:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

Assume that

1. $E(\epsilon \mid \mathbf{X}) = 0$.
2. $\text{var}(\epsilon \mid \mathbf{X}) = \sigma^2\mathbf{I}$, i.e., the errors have constant variance and are uncorrelated.
3. $E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\beta$
4. $\mathbf{X}$ is a full-rank matrix.

Then the **Gauss-Markov** states that the OLS estimator of β,

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^T\mathbf{X}^T\mathbf{y},$$

has the minimum variance among all unbiased estimators of β and this estimator is unique.

Some comments:

- Assumption 3 guarantees that we have hypothesized the correct model, i.e., that we have included exactly the correct regressors in our model. Not only are we fitting a linear model to the data, but our hypothesized model is actually correct.
- Assumption 4 ensures that the OLS estimator can be computed (otherwise, there is no unique solution).
- The Gauss-Markov theorem only applies to unbiased estimators of β. Biased estimators could have a smaller variance.
- The Gauss-Markov theorem states that no unbiased estimator of β can have a smaller variance than $\hat{\beta}$.
- The OLS estimator uniquely has the minimum variance property, meaning that if an $\tilde{\beta}$ is another unbiased estimator of β and $\text{var}(\tilde{\beta}) = \text{var}(\hat{\beta})$, then in fact the two estimators are identical and $\tilde{\beta} = \hat{\beta}$.

We do not prove this theorem.