Violating the homoskedasticity assumption of the linear regression model

Econometrics for minor Finance, Lecture 6

Tunga Kantarcı, Fall 2025

Linear regression model: The homoskedasticity assumption

The population linear regression model with one regressor, for observational unit i, is

$$y_i = \beta x_i + u_i$$

Linear regression model: The homoskedasticity assumption

We assumed that

$$\text{Var}\left[u_i \mid x_i\right] = \sigma^2$$

We called this the homoskedasticity assumption. In this lecture we will study what this assumption means, and study the implications if we violate it.

Linear regression model: The homoskedasticity assumption

The population linear regression model with one regressor is

$$y = \beta x + u$$

Recall from an earlier lecture that if

$$\mathrm{E}[u \mid x] = 0$$

we have

$$\mathrm{E}[y \mid x] = \beta x$$

Plug this back in the model to obtain

$$y = \mathrm{E}[y \mid x] + u$$

y is decomposed into a non-random systematic component and a random component.

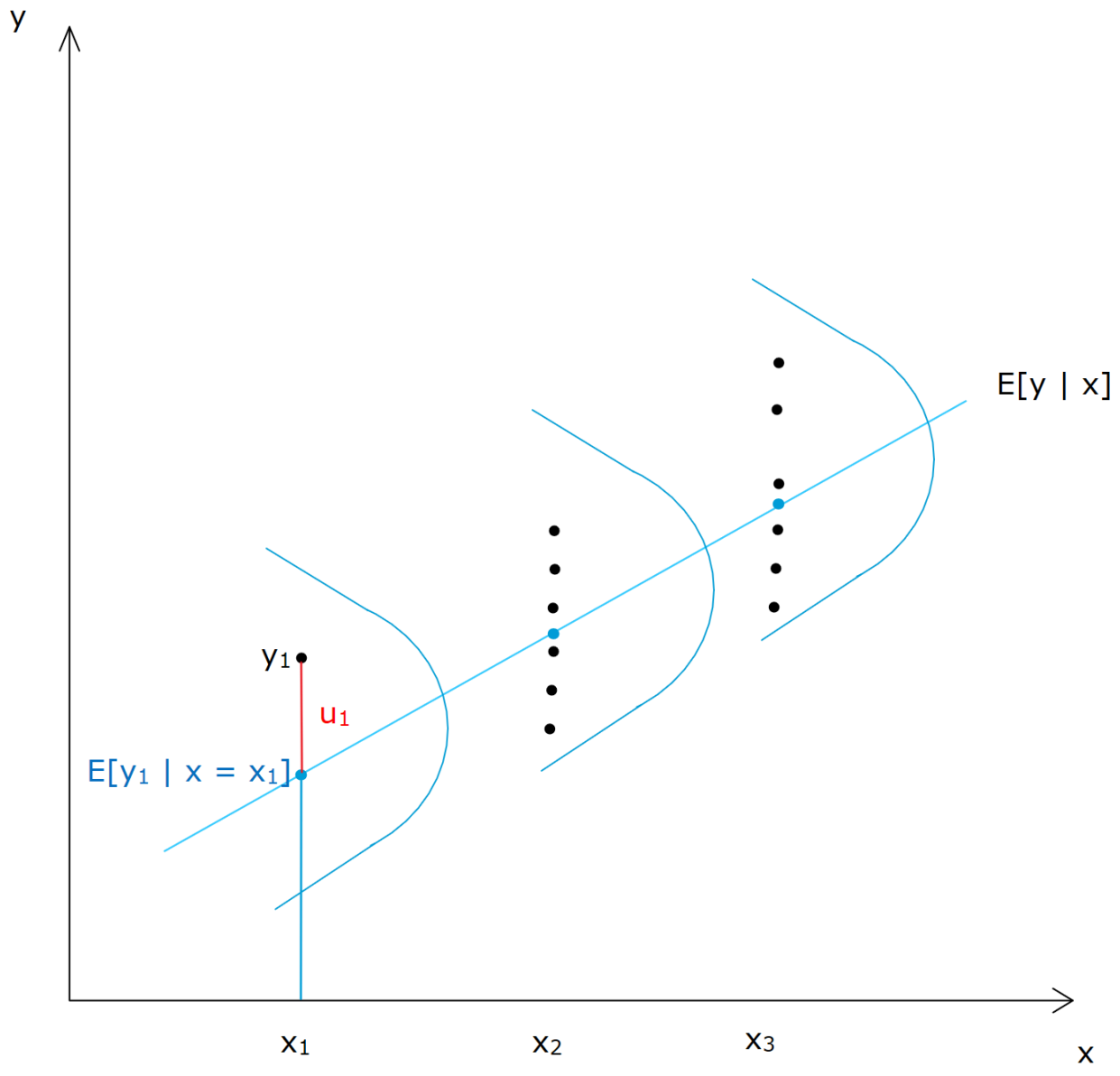
Linear regression model: The homoskedasticity assumption

For observation i we have

$$y_i = \mathrm{E}\left[y_i \mid x_i\right] + u_i$$

The error for i represents dispersion around the conditional expectation function.

Linear regression model: The homoskedasticity assumption



Linear regression model: The homoskedasticity assumption

We summarize dispersion around a mean with variance.

The variance of interest is

$$\text{Var}\left[u_i \mid x_i\right]$$

The question is whether this variance is constant across i. Here across i means across i with different regressor values

$$x_i$$

Linear regression model: The homoskedasticity assumption

The problem is

$$\text{Var}\left[u_i \mid x_i\right]$$

is a population variance so we cannot check whether it is constant across i. But, using sample data, we can approximate

$$u_i$$

with residuals

$$\hat{u}_i$$

and check whether the variance of residuals

$$\text{Var}\left[\hat{u}_i \mid x_i\right]$$

is constant across regressor values

$$x_i$$

Linear regression model: The homoskedasticity assumption

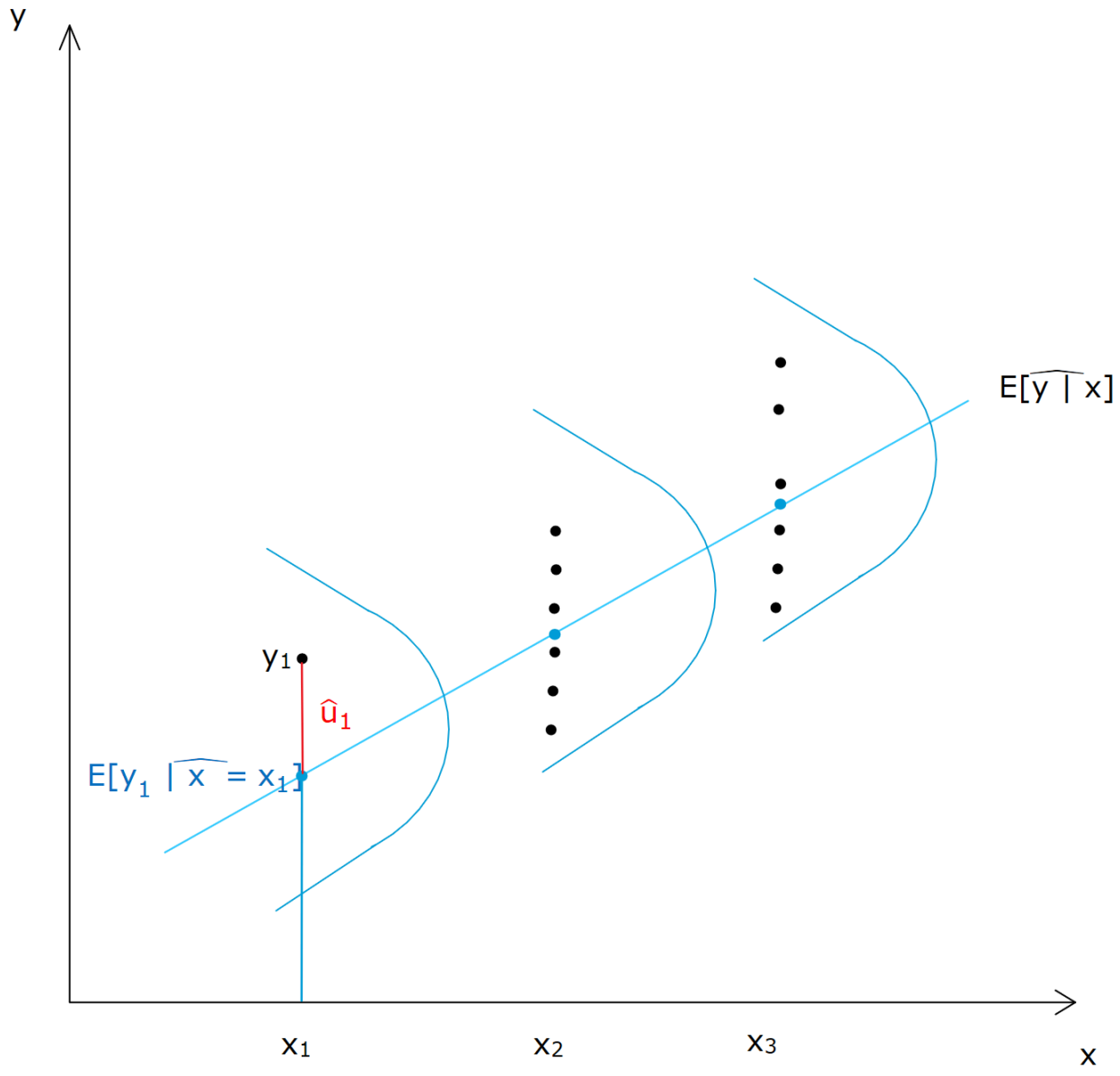
Recall that

$$\hat{u}_i = y_i - \hat{y}_i$$

and that

$$\hat{y}_i = \widehat{\mathsf{E}\left[y_i \mid x_i\right]}$$

Linear regression model: The homoskedasticity assumption



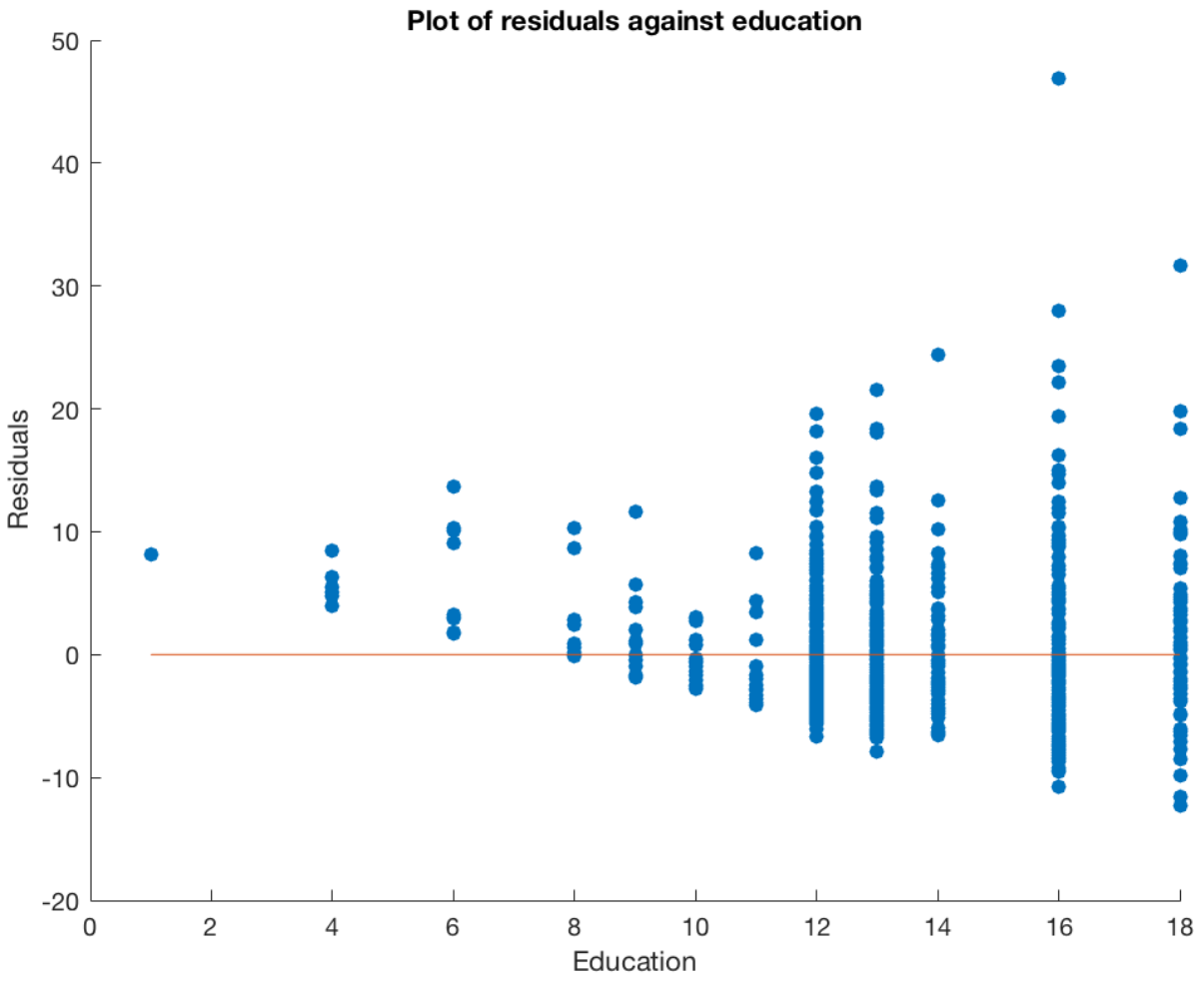
Linear regression model: The homoskedasticity assumption: Example

Suppose we want to explain wage with education using the model

$$wage = \beta_0 + \beta_1 educ + u$$

We can plot the residuals from this regression against education to check whether residuals exhibit constant variance across values of education.

Linear regression model: The homoskedasticity assumption: Example



Linear regression model: The homoskedasticity assumption: Example

Why might the residuals in this model fail to have a constant variance across values of education?

Linear regression model: The homoskedasticity assumption: Example

Think of job opportunities. More education typically means a wider variety of job opportunities.

As a result, wages are more variable at higher levels of education.

But job opportunities are difficult to observe directly, so they enter the error term of the regression model.

If they enter the error, then the errors will show more variation at higher levels of education.

Conclusion is that the variance of residuals $\hat{u}_i$ conditional on x_i is not constant.

Linear regression model: The homoskedasticity assumption

Consider again the linear model

$$y_i = \beta x_i + u_i$$

Take the variance conditional on x_i

$$\begin{aligned} \text{Var}\left[y_i \mid x_i\right] &= \text{Var}\left[\beta x_i \mid x_i\right] + \text{Var}\left[u_i \mid x_i\right] \\ &= \beta^2 \text{Var}\left[x_i \mid x_i\right] + \text{Var}\left[u_i \mid x_i\right] \\ &= \text{Var}\left[u_i \mid x_i\right] \end{aligned}$$

This shows that variance of the error drives the variance of the outcome. This is not surprising. If we condition on x, there is no randomness stemming from x, and error variance will drive the outcome variance.

Linear regression model: The homoskedasticity assumption

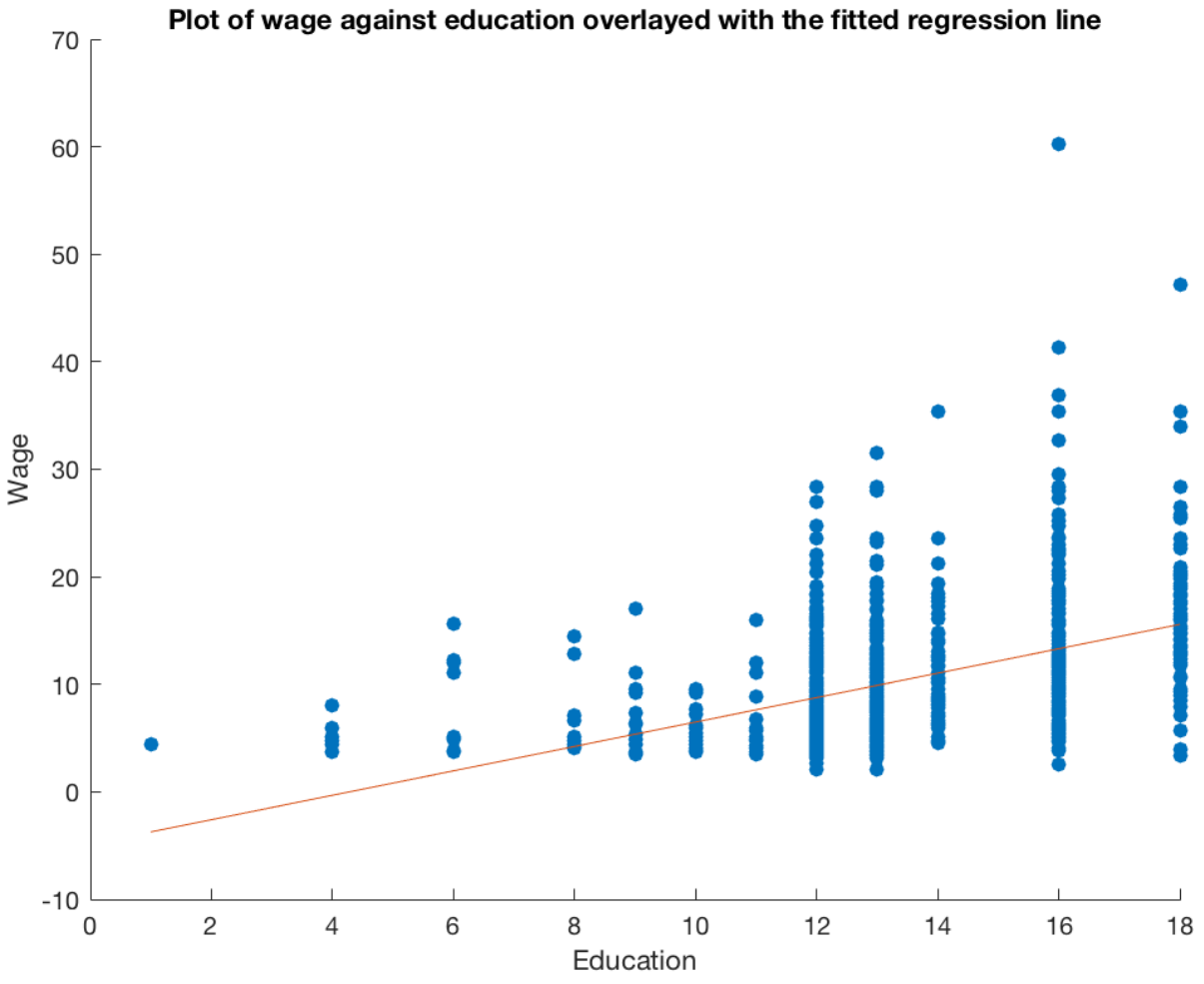
If

$$\text{Var}\left[\hat{u}_i \mid x_i\right] = \hat{\sigma}_i^2$$

That is, homoskedasticity does not hold. This should signal that

$$\text{Var}\left[y_i \mid x_i\right] = \sigma_i^2$$

Linear regression model: The homoskedasticity assumption: Example



Linear regression model: The homoskedasticity assumption

Let's study what we call the variance and covariance matrix of

$$u_i$$

Linear regression model: The homoskedasticity assumption

$$\mathrm{Var}\left[u_i \mid x\right] = \mathrm{E}\left[u_i u_i \mid x\right] - \mathrm{E}\left[u_i \mid x\right]\mathrm{E}\left[u_i \mid x\right]$$

If

$$\mathrm{E}\left[u_i \mid x\right] = 0$$

we have

$$\mathrm{Var}\left[u_i \mid x\right] = \mathrm{E}\left[u_i u_i \mid x\right]$$

If homoskeadasticity holds we have

$$\mathrm{E}\left[u_i u_i \mid x\right] = \sigma^2$$

Linear regression model: The no autocorrelation assumption

$$\mathrm{Cov}\left[u_i, u_j \mid x\right] = \mathrm{E}\left[u_i u_j \mid x\right] - \mathrm{E}\left[u_i \mid x\right]\mathrm{E}\left[u_j \mid x\right]$$

If

$$\mathrm{E}\left[u_i \mid x\right] = 0$$

we have

$$\mathrm{Cov}\left[u_i, u_j \mid x\right] = \mathrm{E}\left[u_i u_j \mid x\right]$$

If no autocorrelation holds, we have

$$\mathrm{E}\left[u_i u_j \mid x\right] = 0$$

Linear regression model: The no autocorrelation assumption

No autocorrelation states that

$$u_i$$

is uncorrelated with every other

$$u_j$$

Linear regression model: The homoskedasticity and no autocorrelation assumption

For observation i, we have

$$\mathrm{E}\left[u_i u_i \mid x\right] = \sigma^2$$

and

$$\mathrm{E}\left[u_i u_j \mid x\right] = 0$$

Linear regression model: The homoskedasticity and no autocorrelation assumption

For n observations, we have a variance and covariance matrix:

$$\begin{aligned}\mathrm{Var}\left[u \mid x\right] &= \mathrm{E}\left[uu' \mid x\right] \\ &= \sigma^2 I_n\end{aligned}$$

where

$$u$$

is $n \times 1$ so

$$uu'$$

is $n \times n$. Hence,

$$\mathrm{E}\left[uu' \mid x\right]$$

is $n \times n$.

Linear regression model: The homoskedasticity and no autocorrelation assumption

$$\mathrm{Var}\left[u \mid x\right] = \mathrm{E}\left[uu' \mid x\right]$$

is called the variance and covariance matrix although we denote it still with

$$\mathrm{Var}$$

This is just convention.

Linear regression model: The homoskedasticity and no autocorrelation assumption

How does

$$\mathrm{E}\left[uu' \mid x\right]$$

look like?

Linear regression model: The homoskedasticity and no autocorrelation assumption

$$
\begin{aligned}
\mathrm{E}\left[uu' \mid x\right] &= \begin{bmatrix} \mathrm{E}\left[u_1u_1 \mid x\right] & \mathrm{E}\left[u_1u_2 \mid x\right] & \dots & \mathrm{E}\left[u_1u_n \mid x\right] \\ \mathrm{E}\left[u_2u_1 \mid x\right] & \mathrm{E}\left[u_2u_2 \mid x\right] & \dots & \mathrm{E}\left[u_2u_n \mid x\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\left[u_nu_1 \mid x\right] & \mathrm{E}\left[u_nu_2 \mid x\right] & \dots & \mathrm{E}\left[u_nu_n \mid x\right] \end{bmatrix} \\
&= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \\
&= \sigma^2 \underbrace{\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}}_{I_n}
\end{aligned}
$$