# The Importance of Model selection

# Why the best model matters

We aim to find the simplest possible model that adequately explains the observed response.

- **over-simplification** risks omitting key predictors leading to incorrect inference ('model mis-specification')
- **over-complexity** may lead to poor predictive behaviour, and weaker (i.e. less powerful) statistical inference.

Consider the model set up

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon = \mathbf{X}\beta + \epsilon = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \epsilon$$

where $\mathbf{X} = \left[\mathbf{X}^{(1)}\ \mathbf{X}^{(2)}\right]$, and where $\beta$, $\beta^{(1)}$ and $\beta^{(2)}$ are $p$, $p - r$ and $r$-dimensional parameter vector and sub-vectors respectively.

# Over-simplification

- **True Model:**

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \epsilon$$

- **Fitted Model:**

$$\mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \epsilon.$$

For the estimators, we have

$$\widehat{\beta}^{(1)} = (\{\mathbf{X}^{(1)}\}^\top\{\mathbf{X}^{(1)}\})^{-1}\{\mathbf{X}^{(1)}\}^\top\mathbf{Y} = \mathbf{A}^{(1)}\mathbf{Y}$$

say, and

$$\widehat{\sigma}^2_{(1)} = \frac{1}{n-(p-r)}\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{H}^{(1)})\mathbf{Y}$$

# Over-simplification (cont.)

The estimator $\widehat{\beta}^{(1)}$ is in general biased:

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}^{(1)}|\mathbf{X}] = \mathbf{A}^{(1)}\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbf{A}^{(1)}(\mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)})$$
$$= \beta^{(1)} + \mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta^{(2)}$$

which does not equal $\beta^{(1)}$ unless

$$\mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta^{(2)} = \mathbf{0}_{p-r};$$

this follows if and only if $\beta^{(2)} = \mathbf{0}_r$ (i.e. the $\mathbf{X}^{(2)}$ predictors are not influential), or

$$\mathbf{A}^{(1)}\mathbf{X}^{(2)} = \mathbf{0}_{p-r,r} \qquad \Longleftrightarrow \qquad \{\mathbf{X}^{(1)}\}^\top\{\mathbf{X}^{(2)}\} = \mathbf{0}_{p-r,r}$$

i.e. the predictors in $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ are **orthogonal**.

# Over-simplification (cont.)

We also have by standard theory for the reduced model

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}^{(1)}|\mathbf{X}] = \sigma^2(\{\mathbf{X}^{(1)}\}^\top\{\mathbf{X}^{(1)}\})^{-1}$$

whereas if the correct full model is fitted, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}|\mathbf{X}] &= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \\ &= \sigma^2\begin{bmatrix} \{\mathbf{X}^{(1)}\}^\top\{\mathbf{X}^{(1)}\} & \{\mathbf{X}^{(1)}\}^\top\{\mathbf{X}^{(2)}\} \\ \{\mathbf{X}^{(2)}\}^\top\{\mathbf{X}^{(1)}\} & \{\mathbf{X}^{(2)}\}^\top\{\mathbf{X}^{(2)}\} \end{bmatrix}^{-1} \\ &= \sigma^2\begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}^{-1}\end{aligned}$$

say. On inverting the block matrix, we have for the variance covariance block for the $\beta^{(1)}$ component

$$\sigma^2(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1}$$

# Over-simplification (cont.)

In general

$$(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1} \geq \mathbf{S}_{11}^{-1}$$

that is

$$(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1} - \mathbf{S}_{11}^{-1}$$

is positive semi-definite.

Thus the variance from the full model is **larger** than that for the reduced model.

# Over-simplification (cont.)

However, recall that the estimator is **biased**; a fairer comparison involves using the **mean-squared error** (MSE) which is the sum

$$\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}^{(1)}|\mathbf{X}] + (\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}^{(1)}|\mathbf{X}] - \beta^{(1)})(\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\beta}^{(1)}|\mathbf{X}] - \beta^{(1)})^{\top}$$

For the full model, the MSE is merely

$$\sigma^2(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1}$$

whereas for the reduced model, the MSE is

$$\sigma^2\mathbf{S}_{11}^{-1} + \mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta^{(2)}\{\beta^{(2)}\}^{\top}\{\mathbf{X}^{(2)}\}^{\top}\{\mathbf{A}^{(1)}\}^{\top}$$

and so which of the two MSEs is larger depends on the magnitude of $\beta^{(2)}$.

# Over-simplification (cont.)

From previous results for quadratic forms, we have that the estimator $\widehat{\sigma}^2_{(1)}$ has expectation

$$\sigma^2 + \frac{1}{n-(p-r)}\{\mathbf{X}^{(2)}\}^\top\{\beta^{(2)}\}^\top(\mathbf{I}_n - \mathbf{H}^{(1)})\mathbf{X}^{(2)}\beta^{(2)}$$

so there is a **positive bias**.

# Over-simplification (cont.)

For prediction at a future value $\mathbf{x}^{\text{new}} = [\mathbf{x}^{\text{new}}_{(1)}, \mathbf{x}^{\text{new}}_{(2)}]$, using the full model we have no prediction bias, and the variance (and MSE) is

$$\sigma^2(1 + \{\mathbf{x}^{\text{new}}\}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{\text{new}}).$$

For the reduced model, the expectation, bias and variance are

$$\text{Expectation} : \mathbf{x}^{\text{new}}\beta^{(1)} + \mathbf{x}^{\text{new}}\mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta^{(2)}$$

$$\text{Bias} : \mathbf{x}^{\text{new}}\mathbf{A}^{(1)}\mathbf{X}^{(2)}\beta^{(2)}$$

$$\text{Variance} : \sigma^2(1 + \{\mathbf{x}^{\text{new}}_{(1)}\}^\top\mathbf{S}_{11}^{-1}\mathbf{x}^{\text{new}}_{(1)})$$

# Over-complexity

- **True Model:**

$$\mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \epsilon;$$

- **Fitted Model:**

$$\mathbf{Y} = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \epsilon.$$

where we know that the true value of $\beta^{(2)} = \mathbf{0}_r$.

Standard theory applies even in this special case; the least squares estimator is unbiased with variance

$$\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2 \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}^{-1}$$

# Over-complexity (cont.)

As before the variance of $\widehat{\beta}^{(1)}$ from the full model is

$$\sigma^2(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})^{-1} \geq \sigma^2\mathbf{S}_{11}^{-1}$$

so the variance is **larger than the model that does not fit spurious variables**.

The estimator of $\sigma^2$ from the full model is unbiased, following results for the sums of squares decomposition and $F$-test.

# Over-complexity (cont.)

For prediction, let $\widehat{\mathbf{Y}}$ be the prediction from the full model, and $\widehat{\mathbf{Y}}^{(1)}$ be the prediction from the correct model. Clearly the predictions $\widehat{\mathbf{Y}}$ are unbiased. Then we may write

$$\widehat{\mathbf{Y}} = (\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}) + \widehat{\mathbf{Y}}^{(1)}$$

and it follows that $(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})$ and $\widehat{\mathbf{Y}}^{(1)}$ are **orthogonal**

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}^{(1)}(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})^\top|\mathbf{X}] = \mathtt{0}_{n,n}$$

(see the Appendix). Thus

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] &= \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})|\mathbf{X}] + \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}] \\ &\geq \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}].\end{aligned}$$

# Over-complexity (cont.)

Thus we conclude that including the spurious variables adversely affects

- the variance of estimators,
- the variance of predictors

In conclusion, we need to guard against omitting important variables, and including spurious variables.

# Appendix
# Prediction and Orthogonality

# Prediction using two blocks of predictors

Suppose that a linear regression model is to be written in terms of two blocks of predictors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$:

$$\mathbf{X} = \left[\mathbf{X}^{(1)}\ \mathbf{X}^{(2)}\right]$$

so that

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon = \mathbf{X}^{(1)}\beta^{(1)} + \mathbf{X}^{(2)}\beta^{(2)} + \epsilon$$

where

$$\beta = \begin{bmatrix} \beta^{(1)} \\ \beta^{(2)} \end{bmatrix}$$

are the parameter vector and sub-vectors.

# Prediction using two blocks of predictors (cont.)

Let

- $\widehat{\mathbf{Y}}$ be the fitted values arising from the fit of the **full** linear regression model
- $\widehat{\mathbf{Y}}^{(1)}$ be the fitted values arising from the fit of the **reduced** linear regression model that presumes $\beta^{(2)} = \mathbf{0}_r$
  - this is the fit after **omitting** the $\mathbf{X}^{(2)}$ predictors.

# Algebraic proof

As

$$\widehat{\mathbf{Y}} = (\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}) + \widehat{\mathbf{Y}}^{(1)}$$

it follows that $(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})$ and $\widehat{\mathbf{Y}}^{(1)}$ are **orthogonal**: to see this, write

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] &= \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})|\mathbf{X}] + \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}] \\ &\quad + 2\mathrm{Cov}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}), \widehat{\mathbf{Y}}^{(1)}|\mathbf{X}]\end{aligned}$$

where

$$\begin{aligned}\mathrm{Cov}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}), \widehat{\mathbf{Y}}^{(1)}|\mathbf{X}] &= \mathrm{Cov}_{\mathbf{Y}|\mathbf{X}}[(\mathbf{H} - \mathbf{H}^{(1)})\mathbf{Y}, \mathbf{H}^{(1)}\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{H} - \mathbf{H}^{(1)})\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}]\{\mathbf{H}^{(1)}\}^{\top} \\ &= \sigma^2(\mathbf{H} - \mathbf{H}^{(1)})\{\mathbf{H}^{(1)}\}^{\top}\end{aligned}$$

# Algebraic proof (cont.)

Recalling that $\mathbf{H}$ and $\mathbf{H}^{(1)}$ are symmetric and idempotent, we have

$$(\mathbf{H} - \mathbf{H}^{(1)})\{\mathbf{H}^{(1)}\}^\top = \mathbf{H}\mathbf{H}^{(1)} - \mathbf{H}^{(1)}.$$

Now $\mathbf{H}$ is a projection matrix mapping points in $\mathbb{R}^n$ onto the space $\mathcal{X}$ spanned by the columns of $\mathbf{X}$; the columns of $\mathbf{H}^{(1)}$ are elements of $\mathbb{R}^n$, but also as

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

the columns of $\mathbf{H}^{(1)}$ are elements of the subspace $\mathcal{X}_1$ spanned by the columns of $\mathbf{X}^{(1)}$. As $\mathcal{X}_1 \subset \mathcal{X}$, we therefore must have

$$\mathbf{H}\mathbf{H}^{(1)} = \mathbf{H}^{(1)}.$$

# Algebraic proof (cont.)

Hence

$$\mathrm{Cov}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}), \widehat{\mathbf{Y}}^{(1)}|\mathbf{X}] = \sigma^2(\mathbf{H} - \mathbf{H}^{(1)})\{\mathbf{H}^{(1)}\}^\top = \mathbf{0}_{n,n}$$

and

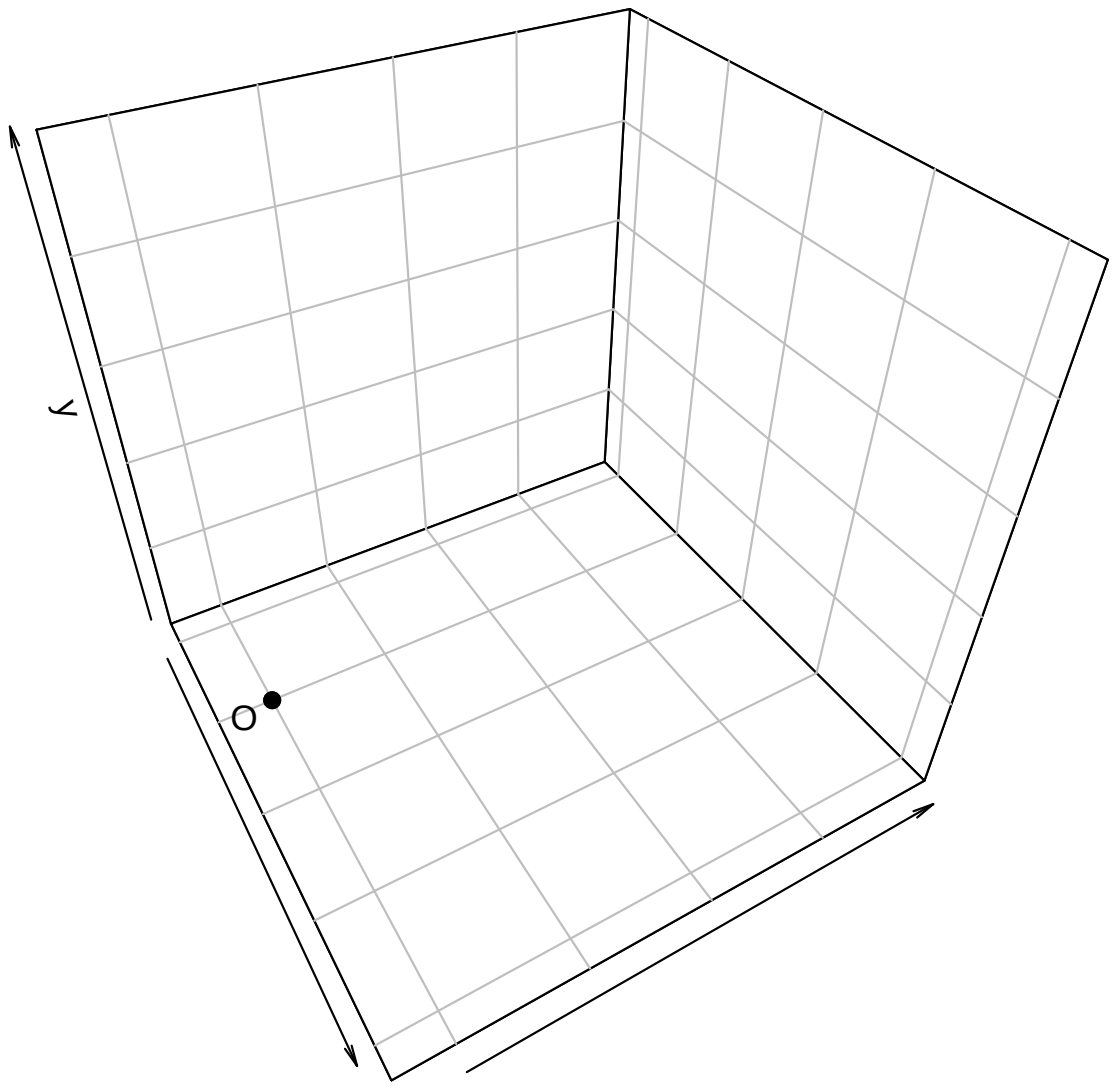
$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] &= \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})|\mathbf{X}] + \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}] \\ &\geq \mathbb{V}\mathrm{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}]\end{aligned}$$

that is, the difference between left hand side and right hand side is positive definite.

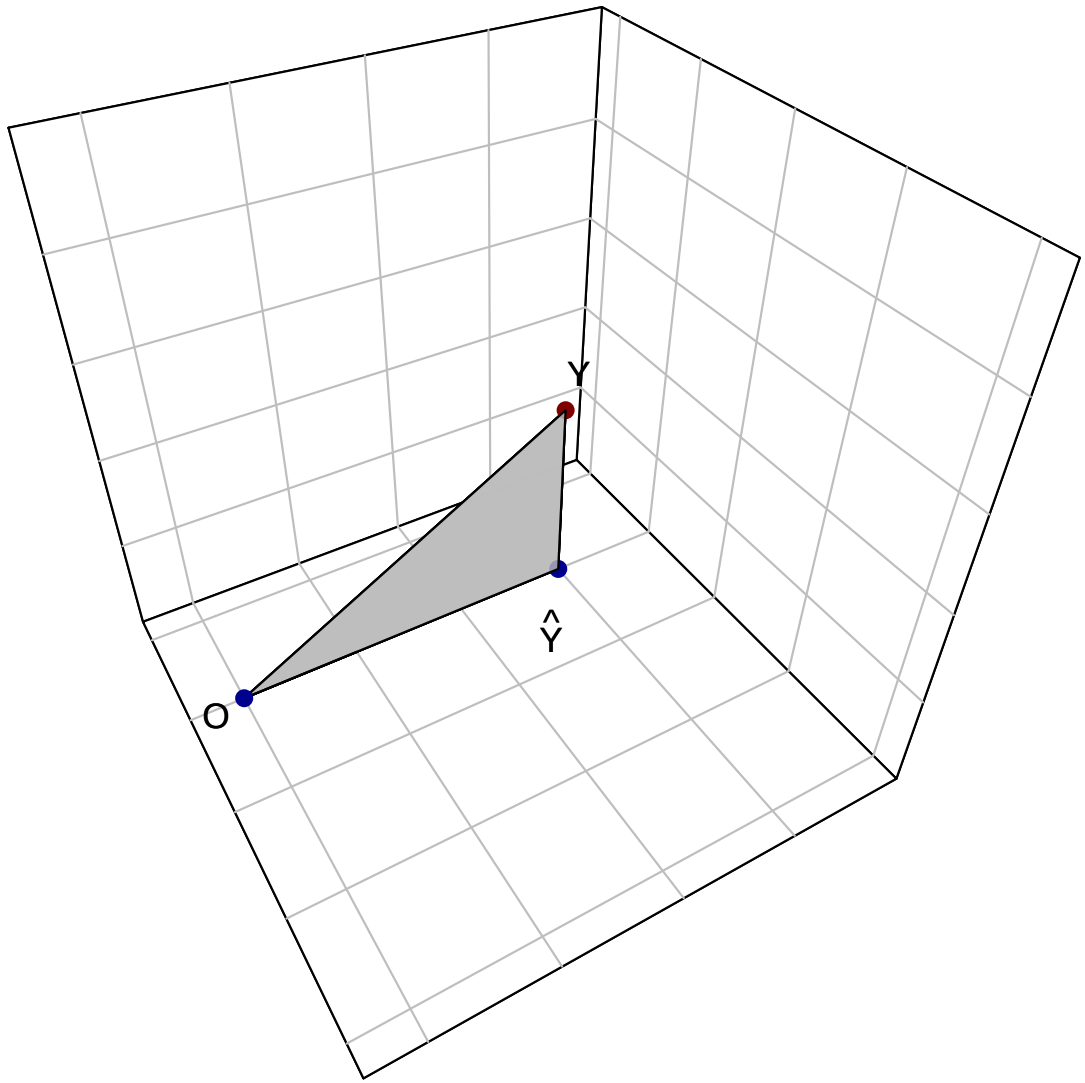
# Geometric proof

A geometric proof follows in a similar fashion; in the following figures we display the data vector in three dimensions, and the model spaces in two and one dimensions for the full model and the true model respectively.

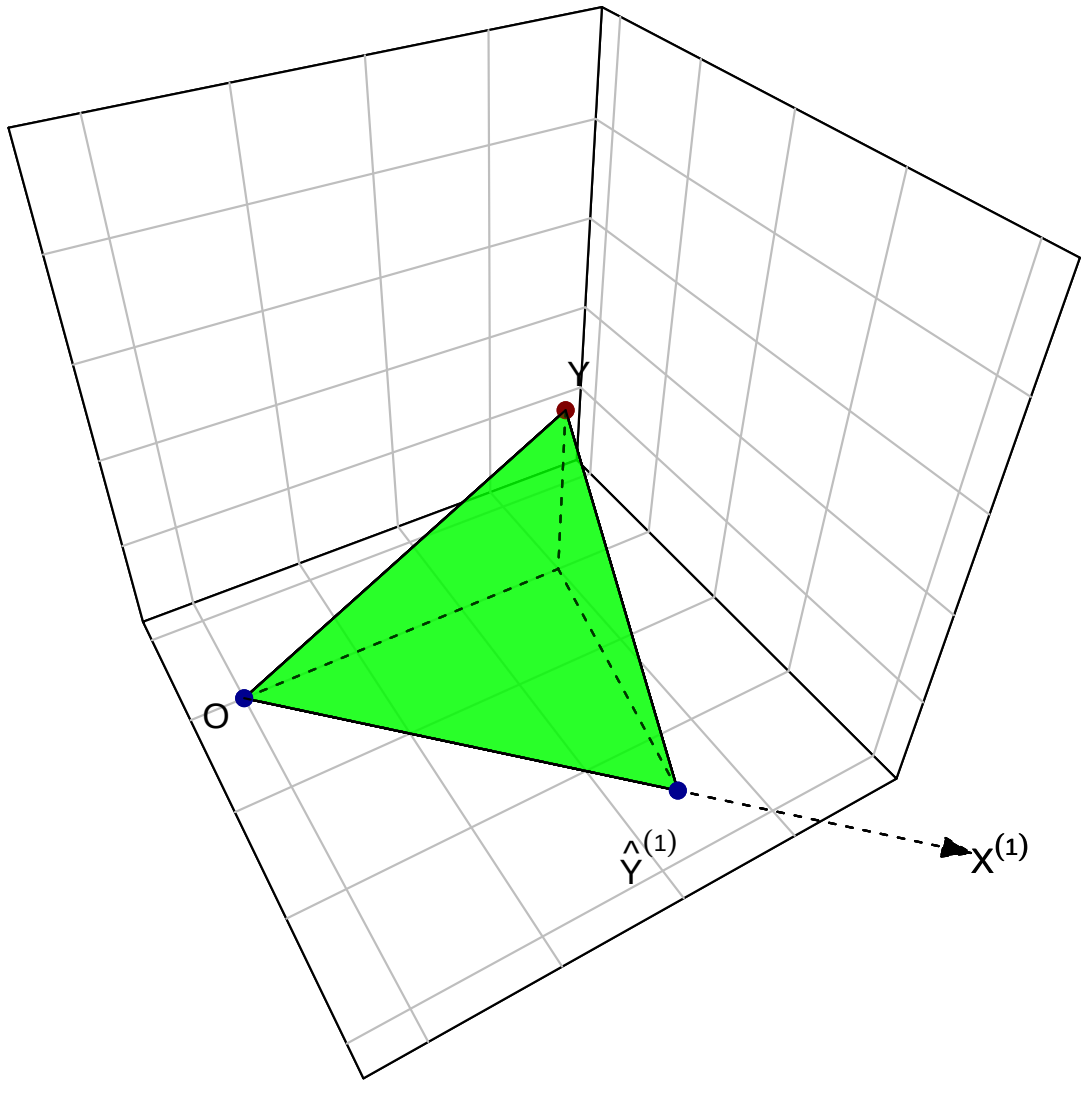
# Observation space (origin marked in $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ plane



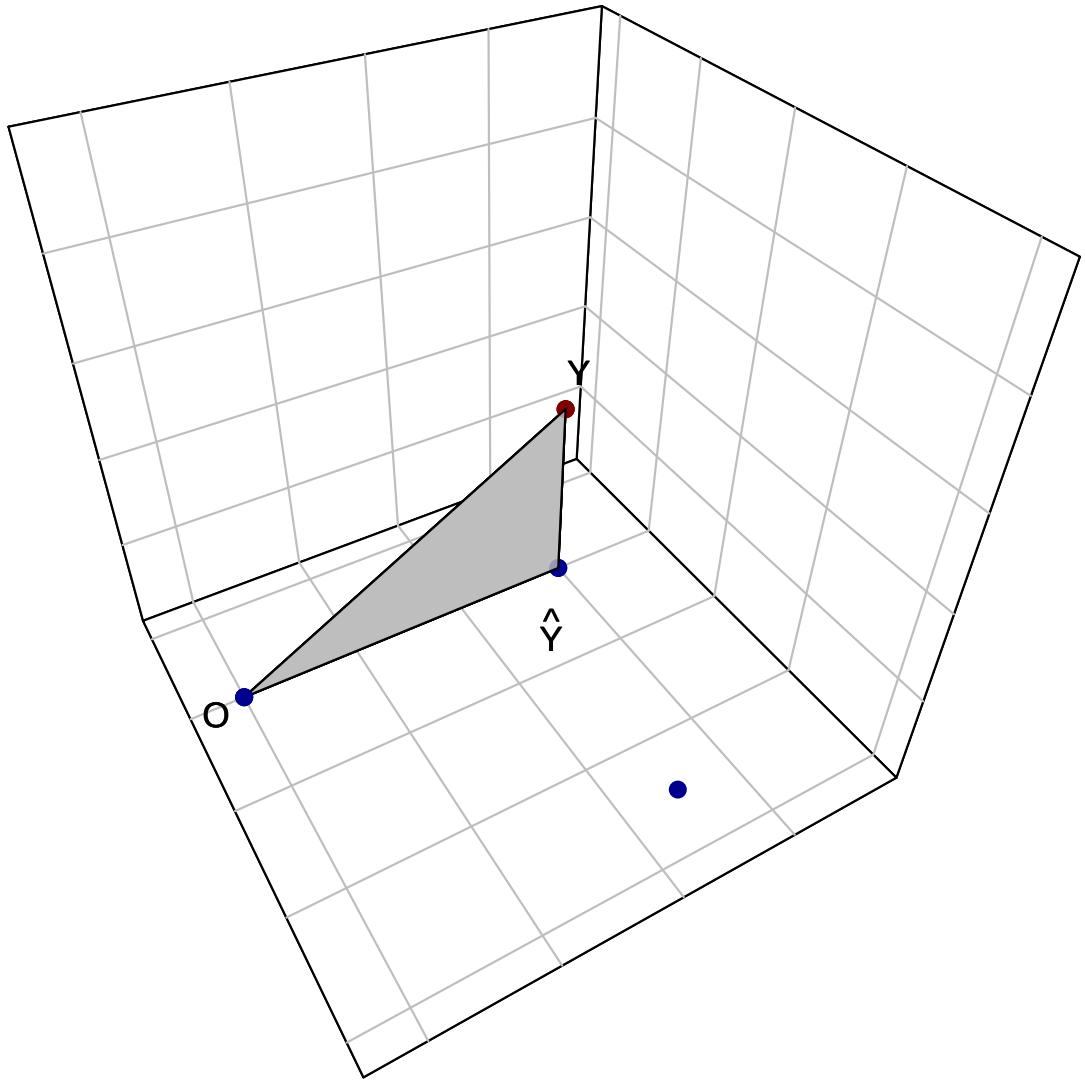
# Fit $\widehat{\mathbf{Y}}$ of response $\mathbf{Y}$: projection onto $(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ plane



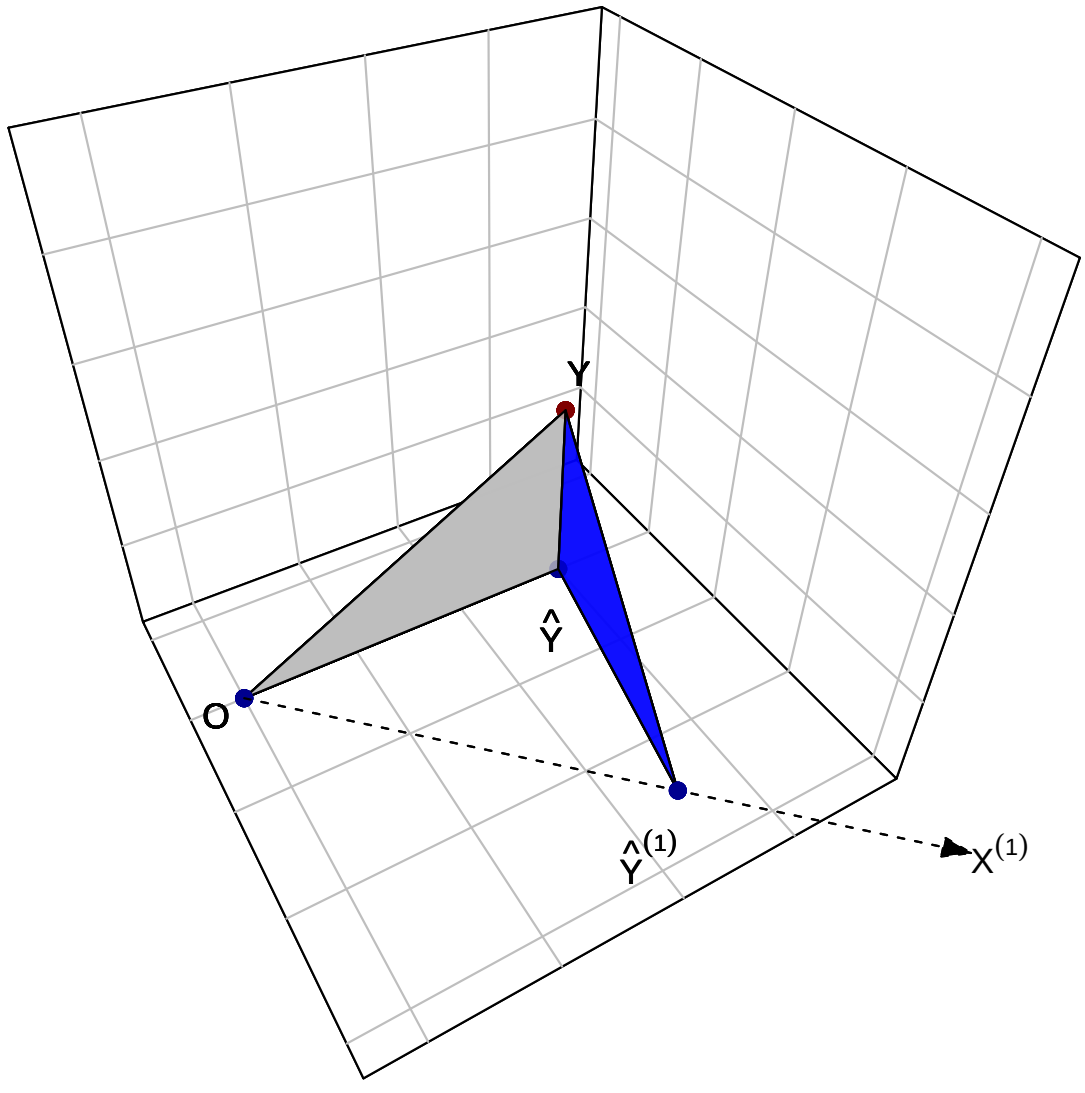
# Fit $\widehat{\mathbf{Y}}^{(1)}$ of response $\mathbf{Y}$: projection onto $\mathbf{X}^{(1)}$ line



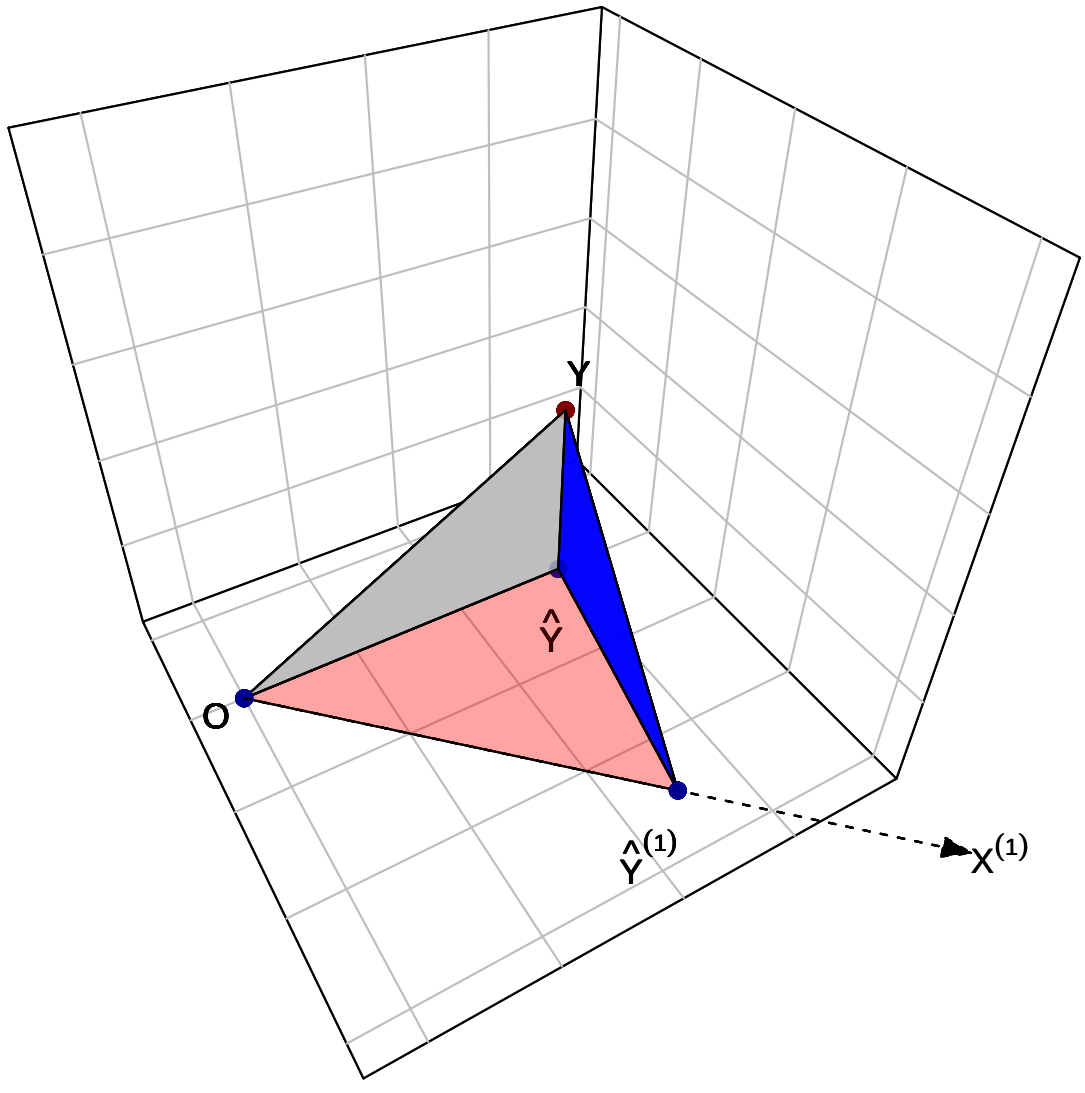
# Fits $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{Y}}^{(1)}$



# Fits $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{Y}}^{(1)}$



# Fits $\widehat{\mathbf{Y}}$ and $\widehat{\mathbf{Y}}^{(1)}$



# Orthogonality

From previous results, we know that after the fit is computed using least squares, we have from the full model

$$\widehat{\mathbf{y}}^\top(\mathbf{y}-\widehat{\mathbf{y}}) = \sum_{i=1}^{n} \widehat{y}_i(y_i-\widehat{y}_i) = 0$$

- see the gray triangle on the previous figure.

Also, we have from the reduced model

$$\{\widehat{\mathbf{y}}^{(1)}\}^\top(\mathbf{y}-\widehat{\mathbf{y}}^{(1)}) = \sum_{i=1}^{n} \widehat{y}_i^{(1)}(y_i-\widehat{y}_i^{(1)}) = 0$$

- see the green triangle on the previous figure.

# Orthogonality and Pythagoras

By these orthogonality results, we know that for the full model

$$
\begin{aligned}
||\mathbf{y}||^2 &= ||\mathbf{y} - \widehat{\mathbf{y}}||^2 + ||\widehat{\mathbf{y}}||^2 \\
\sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} \widehat{y}_i^2 \\
\overline{\mathrm{SS}}_{\mathrm{T}} &= \mathrm{SS}_{\mathrm{Res}} + \overline{\mathrm{SS}}_{\mathrm{R}} \\
c^2 &= b^2 + a^2
\end{aligned}
$$

say.

# Orthogonality and Pythagoras

For the reduced model

$$
\begin{aligned}
||\mathbf{y}||^2 &= ||\mathbf{y} - \widehat{\mathbf{y}}^{(1)}||^2 + ||\widehat{\mathbf{y}}^{(1)}||^2 \\
\sum_{i=1}^{n} y_i^2 &= \sum_{i=1}^{n} (y_i - \widehat{y}_i^{(1)})^2 + \sum_{i=1}^{n} \{\widehat{y}_i^{(1)}\}^2 \\
\overline{\mathrm{SS}}_{\mathrm{T}} &= \mathrm{SS}_{\mathrm{Res}}^{(1)} + \overline{\mathrm{SS}}_{\mathrm{R}}^{(1)} \\
c^2 &= e^2 + d^2
\end{aligned}
$$

say.

# Orthogonality and Pythagoras

Therefore

$$\mathrm{SS}_{\mathrm{Res}} + \overline{\mathrm{SS}}_{\mathrm{R}} = \mathrm{SS}^{(1)}_{\mathrm{Res}} + \overline{\mathrm{SS}}^{(1)}_{\mathrm{R}}$$

or

$$b^2 + a^2 = e^2 + d^2$$

# Orthogonality and Pythagoras

However, the residual vector from the full model, $\mathbf{y} - \widehat{\mathbf{y}}$, has the property that

$$\{\mathbf{X}^{(1)}\}^\top(\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{0}$$

as the columns of $\mathbf{X}^{(1)}$ are used in the fit that produces $\widehat{\mathbf{y}}$.

Therefore the (blue) triangle $Y\widehat{Y}\widehat{Y}^{(1)}$ is a right angle triangle, and we have

$$\begin{aligned}
||\mathbf{y} - \widehat{\mathbf{y}}^{(1)}||^2 &= ||\mathbf{y} - \widehat{\mathbf{y}}||^2 + ||\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(1)}||^2 \\
\sum_{i=1}^{n}(y_i - \widehat{y}_i^{(1)})^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \widehat{y}_i^{(1)})^2 \\
e^2 &= b^2 + f^2
\end{aligned}$$

# Orthogonality and Pythagoras

Thus by the previous result

$$\begin{aligned} b^2 + a^2 &= e^2 + d^2 \\ \implies b^2 + a^2 &= (b^2 + f^2) + d^2 \\ \implies a^2 &= f^2 + d^2 \end{aligned}$$

that is

$$||\widehat{\mathbf{y}}||^2 = ||\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(1)}||^2 + ||\widehat{\mathbf{y}}^{(1)}||^2$$

or equivalently

$$\sum_{i=1}^{n} \widehat{y}_i^2 = \sum_{i=1}^{n} (\widehat{y}_i - \widehat{y}_i^{(1)})^2 + \sum_{i=1}^{n} \{\widehat{y}_i^{(1)}\}^2$$

# Orthogonality and Pythagoras

Hence the vectors

$$(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}^{(1)}) \qquad \text{and} \qquad \widehat{\mathbf{y}}^{(1)}$$

are **orthogonal**, so that

$$\sum_{i=1}^{n} (\widehat{y}_i - \widehat{y}_i^{(1)}) \widehat{y}_i^{(1)} = 0$$

Carrying these results forward to the random variable versions, we have that

$$(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}) \qquad \text{and} \qquad \widehat{\mathbf{Y}}^{(1)}$$

are **uncorrelated**.

# Prediction, Orthogonality and Variances

Suppose now that the true (data-generating) model has $\beta^{(2)} = 0$; that is, the model that uses $\mathbf{X}$ as the predictor matrix is using extra variables that **do not contribute significantly** to the fit. Then we have

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}^{(1)}}[\mathbf{Y}|\mathbf{X}^{(1)}] = \mathbf{X}^{(1)}\beta^{(1)}$$

and also that

$$\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] = \mathbb{E}_{\mathbf{Y}|\mathbf{X}^{(1)}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}^{(1)}] = \mathbf{X}^{(1)}\beta^{(1)}.$$

that is, the predictions are identical (and correct) in expectation under the full and reduced model.

# Prediction, Orthogonality and Variances

However

$$\begin{aligned}\mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[\widehat{\mathbf{Y}}|\mathbf{X}] &= \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}) + \widehat{\mathbf{Y}}^{(1)}|\mathbf{X}] \\ &= \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}}[(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)})|\mathbf{X}] + \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}^{(1)}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}^{(1)}] \\ &\geq \mathbb{V}\text{ar}_{\mathbf{Y}|\mathbf{X}^{(1)}}[\widehat{\mathbf{Y}}^{(1)}|\mathbf{X}^{(1)}]\end{aligned}$$

with the second line following by the uncorrelatedness of

$$(\widehat{\mathbf{Y}} - \widehat{\mathbf{Y}}^{(1)}) \qquad \text{and} \qquad \widehat{\mathbf{Y}}^{(1)}.$$

Thus the variances of predictions under the full model are at least as large as the variances under the reduced model.