STAC67: Regression Analysis

Lecture 11

Thibault Randrianarisoa

October 14, 2026

University of Toronto Scarborough

Review of Lecture 10

The model with $p-1$ predictors, and p parameters:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{X} \text{ of size } n \times p, \qquad E(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \boldsymbol{\sigma^2}\{\boldsymbol{\epsilon}\} = \sigma^2 \boldsymbol{I}.$$

β_k: change in the mean of Y per unit of X_k, the other predictors held fixed.

Least squares, from the normal equations $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$ (one solution when $\boldsymbol{X}$ has full column rank):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \qquad \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{Y}, \qquad \boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \qquad SSE = \boldsymbol{e}'\boldsymbol{e}, \qquad MSE = \frac{SSE}{n-p}.$$

Properties, from the rule $\boldsymbol{\sigma^2}\{\boldsymbol{A}\boldsymbol{Y}\} = \boldsymbol{A}\,\boldsymbol{\sigma^2}\{\boldsymbol{Y}\}\,\boldsymbol{A}'$:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \qquad \boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}\} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \qquad \boldsymbol{s^2}\{\hat{\boldsymbol{\beta}}\} = MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

$E(\hat{\boldsymbol{Y}}) = \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} = \sigma^2\boldsymbol{H}, \qquad$ normal errors: $\hat{\boldsymbol{\beta}} \sim N\big(\boldsymbol{\beta},\, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\big).$

Mean responses and new observations

The mean response at a new point

At predictor values $X_{h1}, \ldots, X_{h,p-1}$, write $\boldsymbol{X}_h = (1, X_{h1}, \ldots, X_{h,p-1})'$. The mean response $E(Y_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$ is estimated by

$$\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}.$$

The mean response at a new point

At predictor values $X_{h1}, \ldots, X_{h,p-1}$, write $\boldsymbol{X}_h = (1, X_{h1}, \ldots, X_{h,p-1})'$. The mean response $E(Y_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$ is estimated by

$$\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}.$$

Again a linear function of $\hat{\boldsymbol{\beta}}$, with the $1 \times p$ constant matrix $\boldsymbol{X}_h'$:

$$E(\hat{Y}_h) = \boldsymbol{X}_h'\boldsymbol{\beta}, \qquad Var(\hat{Y}_h) = \boldsymbol{X}_h'\,\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}\}\,\boldsymbol{X}_h = \sigma^2\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h,$$

estimated by

$$s^2\{\hat{Y}_h\} = MSE\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h.$$

The mean response at a new point

At predictor values $X_{h1}, \ldots, X_{h,p-1}$, write $\boldsymbol{X}_h = (1, X_{h1}, \ldots, X_{h,p-1})'$. The mean response $E(Y_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$ is estimated by

$$\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}.$$

Again a linear function of $\hat{\boldsymbol{\beta}}$, with the $1 \times p$ constant matrix $\boldsymbol{X}_h'$:

$$E(\hat{Y}_h) = \boldsymbol{X}_h'\boldsymbol{\beta}, \qquad Var(\hat{Y}_h) = \boldsymbol{X}_h'\,\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}\}\,\boldsymbol{X}_h = \sigma^2\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h,$$

estimated by

$$s^2\{\hat{Y}_h\} = MSE\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h.$$

With one predictor, this is the formula of Lecture 6: next slide.

For a CI of level $1-\alpha$, take $t(1-\alpha/2; n-p)$ and compute $\hat{E}(\hat{Y}_h) \pm t(1-\alpha/2; n-p)\, s\{\hat{Y}_h\}$

With one predictor: back to Lecture 6

With one predictor, $\boldsymbol{X}_h = (1, X_h)'$ and, from Lecture 9,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} & -\frac{\bar{X}}{S_{xx}} \\ -\frac{\bar{X}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix}.$$

With one predictor: back to Lecture 6

With one predictor, $\boldsymbol{X}_h = (1, X_h)'$ and, from Lecture 9,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} & -\frac{\bar{X}}{S_{xx}} \\ -\frac{\bar{X}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix}.$$

Multiplying out $\begin{bmatrix} 1 & X_h \end{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} \begin{bmatrix} 1 \\ X_h \end{bmatrix}$:

$$\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h = \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} - 2\frac{X_h\bar{X}}{S_{xx}} + \frac{X_h^2}{S_{xx}} = \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}.$$

With one predictor: back to Lecture 6

With one predictor, $\boldsymbol{X}_h = (1, X_h)'$ and, from Lecture 9,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} & -\frac{\bar{X}}{S_{xx}} \\ -\frac{\bar{X}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix}.$$

Multiplying out $\begin{bmatrix} 1 & X_h \end{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} \begin{bmatrix} 1 \\ X_h \end{bmatrix}$:

$$\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h = \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} - 2\,\frac{X_h\bar{X}}{S_{xx}} + \frac{X_h^2}{S_{xx}} = \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}.$$

Hence

$$s^2\{\hat{Y}_h\} = MSE\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{xx}}\right),$$

the formula of Lecture 6: the precision is best at $X_h = \bar{X}$.

Predicting a new observation

A new observation at $\boldsymbol{X}_h$ is $Y_{h(\text{new})} = \boldsymbol{X}_h'\boldsymbol{\beta} + \epsilon_{h(\text{new})}$. Its error is independent of $\epsilon_1, \ldots, \epsilon_n$, so $Y_{h(\text{new})}$ is independent of $\hat{Y}_h$, which depends only on $Y_1, \ldots, Y_n$.

Predicting a new observation

A new observation at $\boldsymbol{X}_h$ is $Y_{h(\text{new})} = \boldsymbol{X}_h'\boldsymbol{\beta} + \epsilon_{h(\text{new})}$. Its error is independent of $\epsilon_1, \ldots, \epsilon_n$, so $Y_{h(\text{new})}$ is independent of $\hat{Y}_h$, which depends only on $Y_1, \ldots, Y_n$.

We predict it by $\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$. The prediction error $Y_{h(\text{new})} - \hat{Y}_h$ has mean

$$E\big(Y_{h(\text{new})} - \hat{Y}_h\big) = \boldsymbol{X}_h'\boldsymbol{\beta} - \boldsymbol{X}_h'\boldsymbol{\beta} = 0.$$

Predicting a new observation

A new observation at $\boldsymbol{X}_h$ is $Y_{h(\text{new})} = \boldsymbol{X}_h'\boldsymbol{\beta} + \epsilon_{h(\text{new})}$. Its error is independent of $\epsilon_1, \ldots, \epsilon_n$, so $Y_{h(\text{new})}$ is independent of $\hat{Y}_h$, which depends only on $Y_1, \ldots, Y_n$.

We predict it by $\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$. The prediction error $Y_{h(\text{new})} - \hat{Y}_h$ has mean

$$E\big(Y_{h(\text{new})} - \hat{Y}_h\big) = \boldsymbol{X}_h'\boldsymbol{\beta} - \boldsymbol{X}_h'\boldsymbol{\beta} = 0.$$

Its variance, using the independence: $Cov\big(Y_{h(new)}, \hat{Y}_h\big) = 0$

$$Var\big(Y_{h(\text{new})} - \hat{Y}_h\big) = \underbrace{Var\big(Y_{h(\text{new})}\big)}_{\sigma^2} + \underbrace{Var(\hat{Y}_h)}_{\sigma^2\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h}$$
$$= \sigma^2\big(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h\big),$$

estimated by $s^2\{\text{pred}\} = MSE\,\big(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h\big)$.

Predicting a new observation

A new observation at $\boldsymbol{X}_h$ is $Y_{h(\text{new})} = \boldsymbol{X}_h'\boldsymbol{\beta} + \epsilon_{h(\text{new})}$. Its error is independent of $\epsilon_1, \ldots, \epsilon_n$, so $Y_{h(\text{new})}$ is independent of $\hat{Y}_h$, which depends only on $Y_1, \ldots, Y_n$.

We predict it by $\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$. The prediction error $Y_{h(\text{new})} - \hat{Y}_h$ has mean

$$E\big(Y_{h(\text{new})} - \hat{Y}_h\big) = \boldsymbol{X}_h'\boldsymbol{\beta} - \boldsymbol{X}_h'\boldsymbol{\beta} = 0.$$

Its variance, using the independence:

$$\begin{aligned} Var\big(Y_{h(\text{new})} - \hat{Y}_h\big) &= \underbrace{Var\big(Y_{h(\text{new})}\big)}_{\sigma^2} + \underbrace{Var(\hat{Y}_h)}_{\sigma^2\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h} \\ &= \sigma^2\big(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h\big), \end{aligned}$$

estimated by $s^2\{\text{pred}\} = MSE\,\big(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h\big)$.

The first σ^2 is the new observation's own error: it does not shrink as n grows.

In R: a car of 3000 lb and 100 hp

vcov(fit2) is $MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and sigma(fit2)^2 is MSE.

```
Xh <- c(1, 3000, 100)
Yh <- sum(Xh * coef(fit2))
s2_Yh <- drop(t(Xh) %*% vcov(fit2) %*% Xh)
c(Yh = Yh, s_Yh = sqrt(s2_Yh),
  s_pred = sqrt(sigma(fit2)^2 + s2_Yh))
```

```
##          Yh        s_Yh      s_pred
## 16.01931725  0.09152615  1.74435944
```

```
new_car <- data.frame(wt = 3000, hp = 100)
predict(fit2, new_car, se.fit = TRUE)$se.fit
```

```
## [1] 0.09152615
```

In R: a car of 3000 lb and 100 hp

`vcov(fit2)` is $MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and `sigma(fit2)^2` is MSE.

```
Xh <- c(1, 3000, 100)
Yh <- sum(Xh * coef(fit2))
s2_Yh <- drop(t(Xh) %*% vcov(fit2) %*% Xh)
c(Yh = Yh, s_Yh = sqrt(s2_Yh),
  s_pred = sqrt(sigma(fit2)^2 + s2_Yh))
```

```
##          Yh        s_Yh      s_pred
## 16.01931725  0.09152615  1.74435944
```

```
new_car <- data.frame(wt = 3000, hp = 100)
predict(fit2, new_car, se.fit = TRUE)$se.fit
```

```
## [1] 0.09152615
```

The mean time to 60 mph of such cars is estimated very precisely ($s\{\hat{Y}_h\} = 0.092$ s), one car's much less ($s\{\text{pred}\} = 1.74$ s).

Summary: estimates and their precision

Estimates	Estimator	Estimated variance
$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$	$MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$
$E(Y_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$	$\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$	$MSE\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h$
$Y_{h(\text{new})}$	$\hat{Y}_h$	$MSE\left(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h\right)$
σ^2	$MSE = SSE/(n-p)$	

Summary: estimates and their precision

Estimates	Estimator	Estimated variance
$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$	$MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$
$E(Y_h) = \boldsymbol{X}_h'\boldsymbol{\beta}$	$\hat{Y}_h = \boldsymbol{X}_h'\hat{\boldsymbol{\beta}}$	$MSE\,\boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h$
$Y_{h(\text{new})}$	$\hat{Y}_h$	$MSE\left(1 + \boldsymbol{X}_h'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_h\right)$
σ^2	$MSE = SSE/(n-p)$	

Next: confidence and prediction intervals, the analysis of variance and R^2 with several predictors.

The geometry of least squares

Data as vectors

Think of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ as one point in the n-dimensional space $\mathbb{R}^n$, and do the same with each column of $\boldsymbol{X}$.

$\boldsymbol{X} \in \mathbb{R}^{n \times p}$

Data as vectors

Think of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ as one point in the n-dimensional space $\mathbb{R}^n$, and do the same with each column of $\boldsymbol{X}$.

Every vector $\boldsymbol{X}\boldsymbol{\beta} = \beta_0\mathbf{1} + \beta_1\boldsymbol{X}_1 + \cdots + \beta_{p-1}\boldsymbol{X}_{p-1}$ is a combination of the columns of $\boldsymbol{X}$. All of them form a p-dimensional subspace, the column space $C(\boldsymbol{X})$, also called the model space.

Data as vectors

Think of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ as one point in the n-dimensional space $\mathbb{R}^n$, and do the same with each column of $\boldsymbol{X}$.

Every vector $\boldsymbol{X}\boldsymbol{\beta} = \beta_0\mathbf{1} + \beta_1\boldsymbol{X}_1 + \cdots + \beta_{p-1}\boldsymbol{X}_{p-1}$ is a combination of the columns of $\boldsymbol{X}$. All of them form a p-dimensional subspace, the column space $C(\boldsymbol{X})$, also called the model space.

The model says that the mean $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ lies in $C(\boldsymbol{X})$. The observed $\boldsymbol{Y}$ is that mean plus noise, and is almost never in $C(\boldsymbol{X})$ itself.

Data as vectors

Think of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$ as one point in the n-dimensional space $\mathbb{R}^n$, and do the same with each column of $\boldsymbol{X}$.

Every vector $\boldsymbol{X}\boldsymbol{\beta} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{X}_1 + \cdots + \beta_{p-1}\boldsymbol{X}_{p-1}$ is a combination of the columns of $\boldsymbol{X}$. All of them form a p-dimensional subspace, the column space $C(\boldsymbol{X})$, also called the model space.

The model says that the mean $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ lies in $C(\boldsymbol{X})$. The observed $\boldsymbol{Y}$ is that mean plus noise, and is almost never in $C(\boldsymbol{X})$ itself.

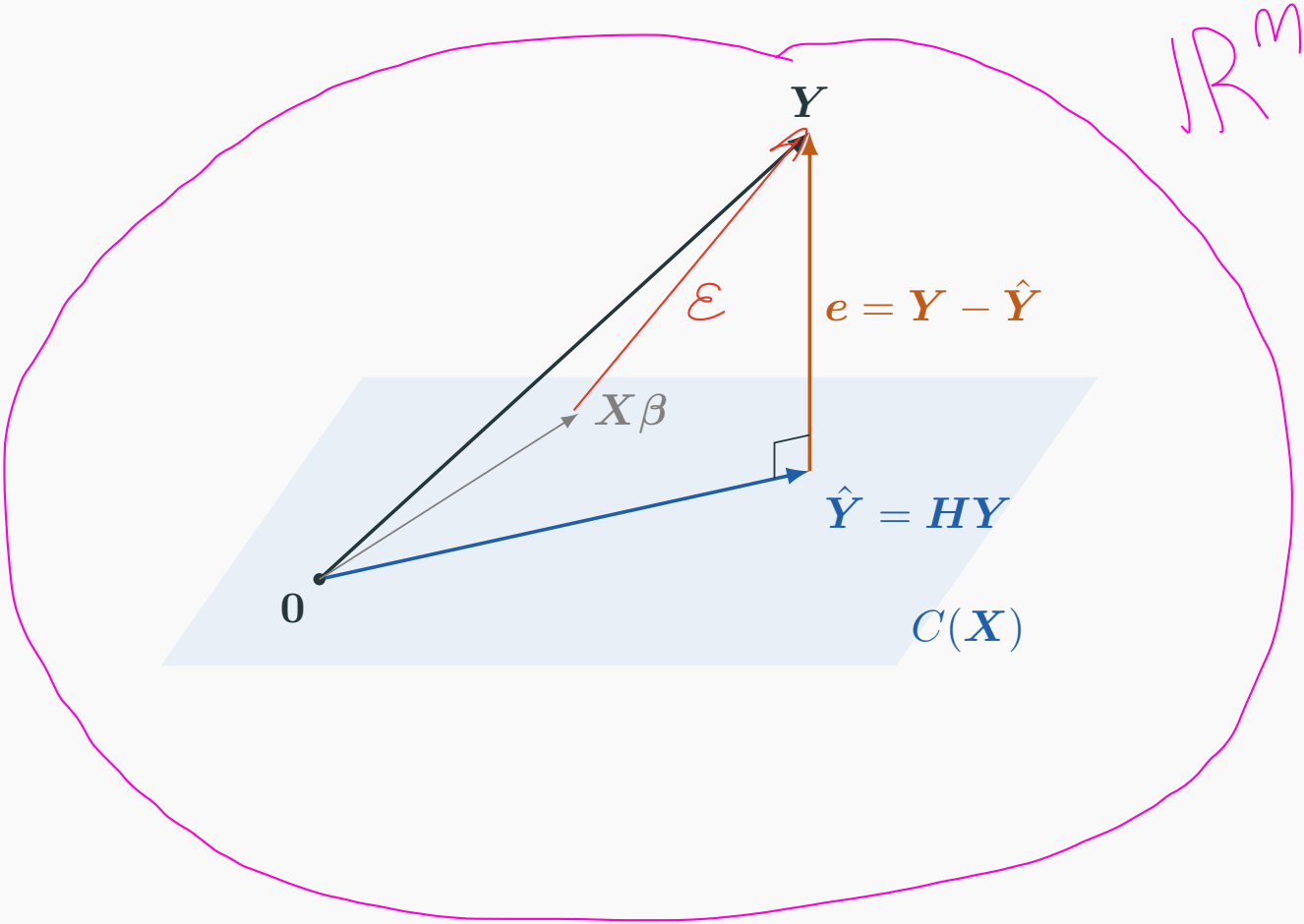
Least squares looks for the point of $C(\boldsymbol{X})$ closest to $\boldsymbol{Y}$:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i)^2 = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2,$$

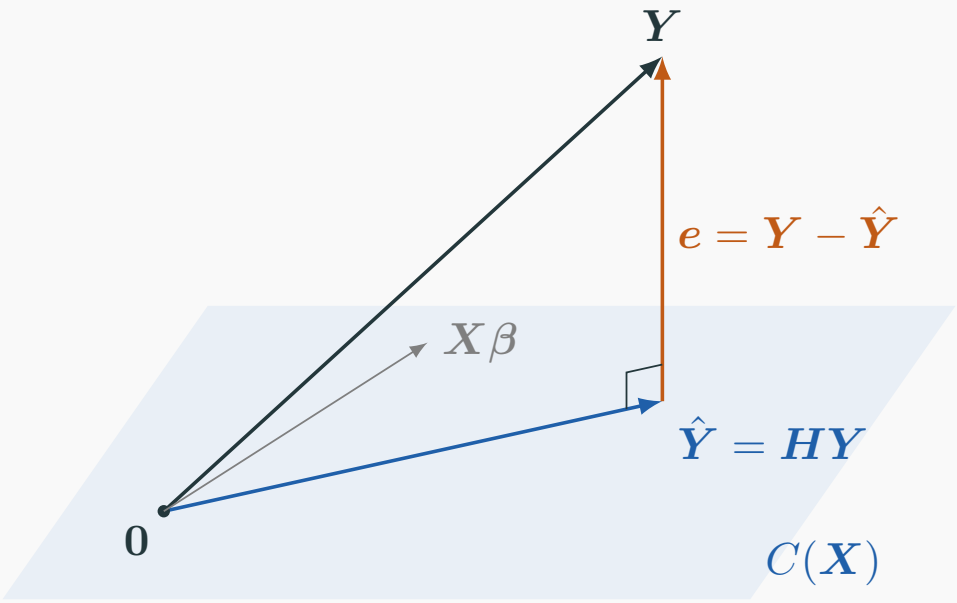
$= (Y - X\beta)^T (Y - X\beta)$

the squared distance between $\boldsymbol{Y}$ and $\boldsymbol{X}\boldsymbol{\beta}$.

Least squares is an orthogonal projection

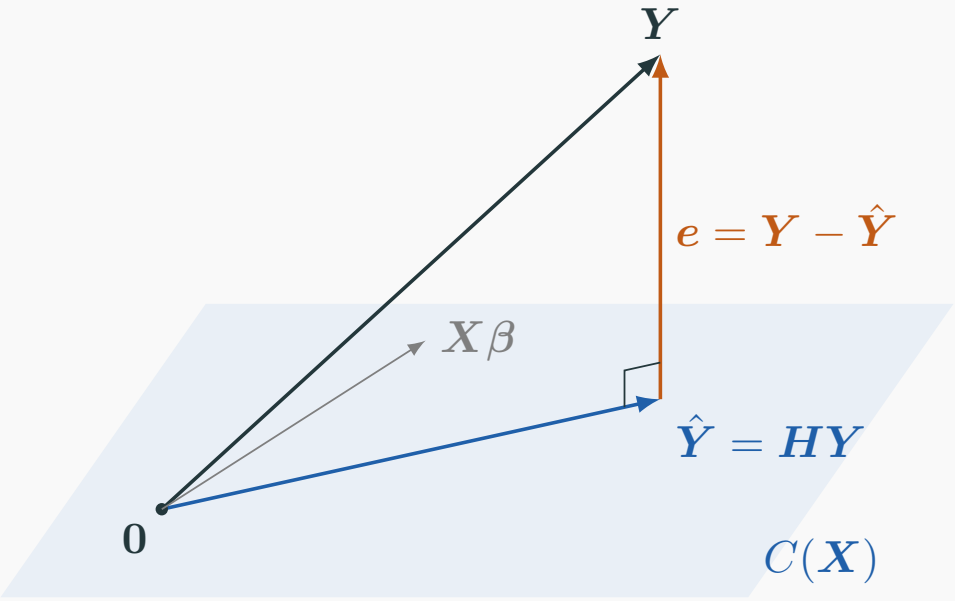


Least squares is an orthogonal projection



The closest point $\hat{\boldsymbol{Y}}$ is the foot of the perpendicular from $\boldsymbol{Y}$: the residual $\boldsymbol{e}$ is orthogonal to every vector of $C(\boldsymbol{X})$.

Least squares is an orthogonal projection



The closest point $\hat{Y}$ is the foot of the perpendicular from Y: the residual e is orthogonal to every vector of $C(X)$.

Any other $X\beta$ in the plane is farther from Y: that is Pythagoras.

$$\|y\|^2 = \|\hat{y}\|^2 + \|e\|^2$$

The normal equations say: orthogonal

e is orthogonal to $C(X)$ exactly when it is orthogonal to each column of X:

$$X'e = 0 \iff X'(Y - X\hat{\beta}) = 0 \iff X'X\hat{\beta} = X'Y.$$

These are the normal equations of Lecture 10.

The normal equations say: orthogonal

$\boldsymbol{e}$ is orthogonal to $C(\boldsymbol{X})$ exactly when it is orthogonal to each column of $\boldsymbol{X}$:

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0} \iff \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0} \iff \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}.$$

These are the normal equations of Lecture 10.

With an intercept, the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, so in particular

$$\boldsymbol{1}'\boldsymbol{e} = \sum_{i=1}^{n} e_i = 0, \qquad \text{and} \qquad \hat{\boldsymbol{Y}}'\boldsymbol{e} = \sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

$\hat{\boldsymbol{Y}} \in C(\boldsymbol{X})$

$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i e_i - \left(\frac{1}{n}\sum e_i\right)\left(\frac{1}{n}\sum \hat{Y}_i\right) = 0$

The sample covariance between $\hat{Y}$ vs. e is 0

The normal equations say: orthogonal

$\boldsymbol{e}$ is orthogonal to $C(\boldsymbol{X})$ exactly when it is orthogonal to each column of $\boldsymbol{X}$:

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}.$$

These are the normal equations of Lecture 10.

With an intercept, the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, so in particular

$$\boldsymbol{1}'\boldsymbol{e} = \sum_{i=1}^{n} e_i = 0, \qquad \text{and} \qquad \hat{\boldsymbol{Y}}'\boldsymbol{e} = \sum_{i=1}^{n} \hat{Y}_i e_i = 0.$$

The second one is why a plot of residuals against fitted values can never show a linear trend (Lecture 8).

The hat matrix is a projection

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ sends any vector to its orthogonal projection on $C(\boldsymbol{X})$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

$= X (X^T X)^{-1} X^T$

The hat matrix is a projection

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ sends any vector to its orthogonal projection on $C(\boldsymbol{X})$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

H is idempotent

- Projecting twice changes nothing: $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. Vectors already in $C(\boldsymbol{X})$ stay put: $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.
- The projection is orthogonal: $\boldsymbol{H}' = \boldsymbol{H}$.
- $\boldsymbol{I} - \boldsymbol{H}$ projects onto the directions orthogonal to $C(\boldsymbol{X})$: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.

$= Y - \hat{Y}$

The hat matrix is a projection

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ sends any vector to its orthogonal projection on $C(\boldsymbol{X})$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

- Projecting twice changes nothing: $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. Vectors already in $C(\boldsymbol{X})$ stay put: $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.
- The projection is orthogonal: $\boldsymbol{H}' = \boldsymbol{H}$.
- $\boldsymbol{I} - \boldsymbol{H}$ projects onto the directions orthogonal to $C(\boldsymbol{X})$: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.

The algebraic proofs are Assignment 2, Question 3.

The hat matrix is a projection

$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ sends any vector to its orthogonal projection on $C(\boldsymbol{X})$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}.$$

- Projecting twice changes nothing: $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. Vectors already in $C(\boldsymbol{X})$ stay put: $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.
- The projection is orthogonal: $\boldsymbol{H}' = \boldsymbol{H}$.
- $\boldsymbol{I} - \boldsymbol{H}$ projects onto the directions orthogonal to $C(\boldsymbol{X})$: $\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.

The algebraic proofs are Assignment 2, Question 3.

The column space has dimension p. The directions orthogonal to it have dimension $n - p$. The residual $\boldsymbol{e}$ belongs to this $(n-p)$-dimensional space.

$\mathbb{R}^n = C(X) \oplus C(X)^\perp$

Properties of the residuals

$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is again a linear function of $\boldsymbol{Y}$. Since $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$:

$$E(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}, \qquad \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2(\boldsymbol{I} - \boldsymbol{H}),$$

and with normal errors, $\boldsymbol{e} \sim N\big(\boldsymbol{0},\ \sigma^2(\boldsymbol{I} - \boldsymbol{H})\big)$.

The proof of $\boldsymbol{\sigma^2}\{\boldsymbol{e}\}$ is Assignment 2, Question 3(c).

Properties of the residuals

$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is again a linear function of $\boldsymbol{Y}$. Since $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$:

$$E(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}, \qquad \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2(\boldsymbol{I} - \boldsymbol{H}),$$

and with normal errors, $\boldsymbol{e} \sim N\big(\boldsymbol{0},\ \sigma^2(\boldsymbol{I} - \boldsymbol{H})\big)$.

The proof of $\boldsymbol{\sigma^2}\{e\}$ is Assignment 2, Question 3(c).

Write h_{ii} for the diagonal entries of $\boldsymbol{H}$. On the diagonals of $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} = \sigma^2\boldsymbol{H}$ and of $\boldsymbol{\sigma^2}\{\boldsymbol{e}\}$:

$$Var(\hat{Y}_i) = h_{ii}\,\sigma^2 \ \geq\ 0, \qquad Var(e_i) = (1 - h_{ii})\,\sigma^2 \ \geq\ 0.$$

Properties of the residuals

$\boldsymbol{e} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ is again a linear function of $\boldsymbol{Y}$. Since $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$:

$$E(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}, \qquad \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2(\boldsymbol{I} - \boldsymbol{H}),$$

and with normal errors, $\boldsymbol{e} \sim N\big(\boldsymbol{0},\ \sigma^2(\boldsymbol{I} - \boldsymbol{H})\big)$.

The proof of $\boldsymbol{\sigma^2}\{e\}$ is Assignment 2, Question 3(c).

Write h_{ii} for the diagonal entries of $\boldsymbol{H}$. On the diagonals of $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} = \sigma^2\boldsymbol{H}$ and of $\boldsymbol{\sigma^2}\{\boldsymbol{e}\}$:

$$Var(\hat{Y}_i) = h_{ii}\,\sigma^2 \ \geq\ 0, \qquad Var(e_i) = (1 - h_{ii})\,\sigma^2 \ \geq\ 0.$$

Hence $0 \leq h_{ii} \leq 1$ for every i.

Fitted values and residuals share the variance

Adding the two variance-covariance matrices:

$$\boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} + \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2\boldsymbol{H} + \sigma^2(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2\boldsymbol{I} = \boldsymbol{\sigma^2}\{\boldsymbol{Y}\}.$$

Fitted values and residuals share the variance

Adding the two variance-covariance matrices:

$$\boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} + \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2\boldsymbol{H} + \sigma^2(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2\boldsymbol{I} = \boldsymbol{\sigma^2}\{\boldsymbol{Y}\}.$$

For each observation:

$$\underbrace{Var(Y_i)}_{\sigma^2} = \underbrace{Var(\hat{Y}_i)}_{h_{ii}\,\sigma^2} + \underbrace{Var(e_i)}_{(1-h_{ii})\,\sigma^2}.$$

Fitted values and residuals share the variance

Adding the two variance-covariance matrices:

$$\boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} + \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2 \boldsymbol{H} + \sigma^2(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2 \boldsymbol{I} = \boldsymbol{\sigma^2}\{\boldsymbol{Y}\}.$$

For each observation:

$$\underbrace{Var(Y_i)}_{\sigma^2} = \underbrace{Var(\hat{Y}_i)}_{h_{ii}\,\sigma^2} + \underbrace{Var(e_i)}_{(1-h_{ii})\,\sigma^2}.$$

Since $\hat{Y}_i = \sum_j h_{ij} Y_j$, h_{ii} is the weight of Y_i in its own fitted value.

- With a large h_{ii}, the fit is pulled towards Y_i: $\hat{Y}_i$ varies a lot and e_i little.
- With a small h_{ii}, the opposite.

Fitted values and residuals share the variance

Adding the two variance-covariance matrices:

$$\boldsymbol{\sigma^2}\{\hat{\boldsymbol{Y}}\} + \boldsymbol{\sigma^2}\{\boldsymbol{e}\} = \sigma^2 \boldsymbol{H} + \sigma^2(\boldsymbol{I} - \boldsymbol{H}) = \sigma^2 \boldsymbol{I} = \boldsymbol{\sigma^2}\{\boldsymbol{Y}\}.$$

For each observation:

$$\underbrace{Var(Y_i)}_{\sigma^2} = \underbrace{Var(\hat{Y}_i)}_{h_{ii}\,\sigma^2} + \underbrace{Var(e_i)}_{(1-h_{ii})\,\sigma^2}.$$

Since $\hat{Y}_i = \sum_j h_{ij} Y_j$, h_{ii} is the weight of Y_i in its own fitted value.

- With a large h_{ii}, the fit is pulled towards Y_i: $\hat{Y}_i$ varies a lot and e_i little.
- With a small h_{ii}, the opposite.

The residuals do not all have the same variance, even when the errors do.

Exercise: a model with only an intercept

Consider the model

$$Y_i = \beta_0 + \epsilon_i, \qquad i = 1, \ldots, n.$$

1. What is the design matrix $\boldsymbol{X}$?
2. What is the value of h_{ii}, for any i?
3. What is the variance of $\hat{Y}_i$?
4. This is a well-known formula for the variance of an estimator. Which one?
5. Is this result surprising? Explain.

Exercise: solution

1. $\boldsymbol{X} = \mathbf{1}$, the column of n ones, so $\boldsymbol{X}'\boldsymbol{X} = \mathbf{1}'\mathbf{1} = n$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = 1/n$.

Exercise: solution

1. $\boldsymbol{X} = \mathbf{1}$, the column of n ones, so $\boldsymbol{X}'\boldsymbol{X} = \mathbf{1}'\mathbf{1} = n$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = 1/n$.

2. $\boldsymbol{H} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}' = \frac{1}{n}\boldsymbol{J}$, where $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\boldsymbol{H}$ is $1/n$: $h_{ii} = 1/n$.

Exercise: solution

1. $\boldsymbol{X} = \mathbf{1}$, the column of n ones, so $\boldsymbol{X}'\boldsymbol{X} = \mathbf{1}'\mathbf{1} = n$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = 1/n$.

2. $\boldsymbol{H} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}' = \frac{1}{n}\boldsymbol{J}$, where $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\boldsymbol{H}$ is $1/n$: $h_{ii} = 1/n$.

3. $Var(\hat{Y}_i) = h_{ii}\,\sigma^2 = \sigma^2/n$.

Exercise: solution

1. $\boldsymbol{X} = \mathbf{1}$, the column of n ones, so $\boldsymbol{X}'\boldsymbol{X} = \mathbf{1}'\mathbf{1} = n$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = 1/n$.

2. $\boldsymbol{H} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}' = \frac{1}{n}\boldsymbol{J}$, where $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\boldsymbol{H}$ is $1/n$: $h_{ii} = 1/n$.

3. $Var(\hat{Y}_i) = h_{ii}\,\sigma^2 = \sigma^2/n$.

4. It is $Var(\bar{Y}) = \sigma^2/n$: indeed $\hat{\boldsymbol{Y}} = \frac{1}{n}\boldsymbol{J}\boldsymbol{Y} = \bar{Y}\mathbf{1}$, so $\hat{Y}_i = \bar{Y}$ for every i.

Exercise: solution

1. $\boldsymbol{X} = \mathbf{1}$, the column of n ones, so $\boldsymbol{X}'\boldsymbol{X} = \mathbf{1}'\mathbf{1} = n$ and $(\boldsymbol{X}'\boldsymbol{X})^{-1} = 1/n$.

2. $\boldsymbol{H} = \mathbf{1}\,\frac{1}{n}\,\mathbf{1}' = \frac{1}{n}\boldsymbol{J}$, where $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\boldsymbol{H}$ is $1/n$: $h_{ii} = 1/n$.

3. $Var(\hat{Y}_i) = h_{ii}\,\sigma^2 = \sigma^2/n$.

4. It is $Var(\bar{Y}) = \sigma^2/n$: indeed $\hat{\boldsymbol{Y}} = \frac{1}{n}\boldsymbol{J}\boldsymbol{Y} = \bar{Y}\mathbf{1}$, so $\hat{Y}_i = \bar{Y}$ for every i.

5. Not surprising: least squares minimizes $\sum_i (Y_i - \beta_0)^2$, and setting the derivative $-2\sum_i (Y_i - \beta_0)$ to zero gives $\hat{\beta}_0 = \bar{Y}$. With no predictor, every fitted value is the sample mean.

Pythagoras and the analysis of variance

Since $\mathbf{1}$ is a column of $\boldsymbol{X}$, the vector $\bar{Y}\mathbf{1}$ is in $C(\boldsymbol{X})$, and so is $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$. Then

$$\boldsymbol{Y} - \bar{Y}\mathbf{1} = (\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}) + \boldsymbol{e}, \qquad \text{with} \quad (\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}) \perp \boldsymbol{e}.$$

Pythagoras and the analysis of variance

Since $\mathbf{1}$ is a column of $\boldsymbol{X}$, the vector $\bar{Y}\mathbf{1}$ is in $C(\boldsymbol{X})$, and so is $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$. Then

$$\boldsymbol{Y} - \bar{Y}\mathbf{1} = (\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}) + \boldsymbol{e}, \qquad \text{with} \quad (\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}) \perp \boldsymbol{e}.$$

Pythagoras:

$$\underbrace{\|\boldsymbol{Y} - \bar{Y}\mathbf{1}\|^2}_{SST} = \underbrace{\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\|^2}_{SSR} + \underbrace{\|\boldsymbol{e}\|^2}_{SSE}.$$

This is the decomposition of Lecture 6, now for any number of predictors.

Pythagoras and the analysis of variance

Since $\mathbf{1}$ is a column of $\boldsymbol{X}$, the vector $\bar{Y}\mathbf{1}$ is in $C(\boldsymbol{X})$, and so is $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$. Then

$$\boldsymbol{Y} - \bar{Y}\mathbf{1} = (\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}) + \boldsymbol{e}, \qquad \text{with} \quad (\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}) \perp \boldsymbol{e}.$$

Pythagoras:

$$\underbrace{\|\boldsymbol{Y} - \bar{Y}\mathbf{1}\|^2}_{SST} = \underbrace{\|\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}\|^2}_{SSR} + \underbrace{\|\boldsymbol{e}\|^2}_{SSE}.$$

This is the decomposition of Lecture 6, now for any number of predictors.

As in Lecture 6, the coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

is the proportion of the variation of Y explained by the regression, now with any number of predictors.

Degrees of freedom are dimensions

Each sum of squares is the squared length of a vector that lives in a subspace:

Vector	lives in a space of dimension	degrees of freedom
$\boldsymbol{Y} - \bar{Y}\mathbf{1}$	$n-1$ (orthogonal to $\mathbf{1}$)	$n-1$ for SST
$\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$	$p-1$ ($C(\boldsymbol{X})$, orthogonal to $\mathbf{1}$)	$p-1$ for SSR
$\boldsymbol{e}$	$n-p$ (orthogonal to $C(\boldsymbol{X})$)	$n-p$ for SSE

Degrees of freedom are dimensions

Each sum of squares is the squared length of a vector that lives in a subspace:

Vector	lives in a space of dimension	degrees of freedom
$\boldsymbol{Y} - \bar{Y}\mathbf{1}$	$n-1$ (orthogonal to $\mathbf{1}$)	$n-1$ for SST
$\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$	$p-1$ ($C(\boldsymbol{X})$, orthogonal to $\mathbf{1}$)	$p-1$ for SSR
$\boldsymbol{e}$	$n-p$ (orthogonal to $C(\boldsymbol{X})$)	$n-p$ for SSE

Why orthogonal to $\mathbf{1}$: with $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$, $\bar{Y}\mathbf{1} = \frac{1}{n}\boldsymbol{J}\boldsymbol{Y}$, so $\boldsymbol{Y} - \bar{Y}\mathbf{1} = (\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$, and

$$\mathbf{1}'\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right) = \mathbf{1}' - \frac{1}{n}(\mathbf{1}'\mathbf{1})\,\mathbf{1}' = \mathbf{1}' - \mathbf{1}' = \mathbf{0}'.$$

Likewise for $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$, since $\mathbf{1}'\hat{\boldsymbol{Y}} = \mathbf{1}'\boldsymbol{Y}$ (because $\mathbf{1}'\boldsymbol{e} = 0$).

Degrees of freedom are dimensions

Each sum of squares is the squared length of a vector that lives in a subspace:

Vector	lives in a space of dimension	degrees of freedom
$\boldsymbol{Y} - \bar{Y}\mathbf{1}$	$n-1$ (orthogonal to $\mathbf{1}$)	$n-1$ for SST
$\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$	$p-1$ ($C(\boldsymbol{X})$, orthogonal to $\mathbf{1}$)	$p-1$ for SSR
$\boldsymbol{e}$	$n-p$ (orthogonal to $C(\boldsymbol{X})$)	$n-p$ for SSE

Why orthogonal to $\mathbf{1}$: with $\boldsymbol{J} = \mathbf{1}\mathbf{1}'$, $\bar{Y}\mathbf{1} = \frac{1}{n}\boldsymbol{J}\boldsymbol{Y}$, so $\boldsymbol{Y} - \bar{Y}\mathbf{1} = (\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$, and

$$\mathbf{1}'\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right) = \mathbf{1}' - \frac{1}{n}(\mathbf{1}'\mathbf{1})\,\mathbf{1}' = \mathbf{1}' - \mathbf{1}' = \mathbf{0}'.$$

Likewise for $\hat{\boldsymbol{Y}} - \bar{Y}\mathbf{1}$, since $\mathbf{1}'\hat{\boldsymbol{Y}} = \mathbf{1}'\boldsymbol{Y}$ (because $\mathbf{1}'\boldsymbol{e} = 0$).

The degrees of freedom add up as the dimensions do: $n - 1 = (p-1) + (n-p)$. With one predictor, $p = 2$: as in Lecture 6.

In R: the acceleration data

```r
X <- model.matrix(fit2)
e <- resid(fit2); Yhat <- fitted(fit2)
round(c(t(X) %*% e, sum(Yhat * e)), 8)    # all zero
```

```
## [1] 0 0 0 0
```

```r
Y <- autompg$acc
c(SST = sum((Y - mean(Y))^2),
  SSR_plus_SSE = sum((Yhat - mean(Y))^2) + sum(e^2))
```

```
##          SST SSR_plus_SSE
##     2969.013     2969.013
```

In R: the acceleration data

```r
X <- model.matrix(fit2)
e <- resid(fit2); Yhat <- fitted(fit2)
round(c(t(X) %*% e, sum(Yhat * e)), 8)    # all zero
```

```
## [1] 0 0 0 0
```

```r
Y <- autompg$acc
c(SST = sum((Y - mean(Y))^2),
  SSR_plus_SSE = sum((Yhat - mean(Y))^2) + sum(e^2))
```

```
##          SST SSR_plus_SSE
##     2969.013     2969.013
```

Next: the distribution of these sums of squares, and the tests they lead to.