

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:

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

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

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

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

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

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

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

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

Mean responses and new observations

The mean response at a new point

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

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

The mean response at a new point

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

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

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

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

estimated by

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

The mean response at a new point

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

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

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

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

estimated by

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

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

With one predictor: back to Lecture 6

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

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

With one predictor: back to Lecture 6

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

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

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

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

With one predictor: back to Lecture 6

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

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

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

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

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

Predicting a new observation

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

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

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

Predicting a new observation

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

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

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

Its variance, using the independence:

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

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

Predicting a new observation

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

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

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

Its variance, using the independence:

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

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

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 $MSE(X'X)^{-1}$ 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 $MSE(X'X)^{-1}$ 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
β	$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$	$MSE (\mathbf{X}'\mathbf{X})^{-1}$
$E(Y_h) = \mathbf{X}_h'\beta$	$\hat{Y}_h = \mathbf{X}_h'\hat{\beta}$	$MSE \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$
$Y_{h(\text{new})}$	$\hat{Y}_h$	$MSE (1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$
σ^2	$MSE = SSE/(n - p)$	

Summary: estimates and their precision

Estimates	Estimator	Estimated variance
β	$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$	$MSE (\mathbf{X}'\mathbf{X})^{-1}$
$E(Y_h) = \mathbf{X}_h'\beta$	$\hat{Y}_h = \mathbf{X}_h'\hat{\beta}$	$MSE \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$
$Y_{h(\text{new})}$	$\hat{Y}_h$	$MSE (1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$
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 $\mathbf{Y} = (Y_1, \dots, Y_n)'$ as **one point** in the n -dimensional space $\mathbb{R}^n$, and do the same with each column of $\mathbf{X}$.

Data as vectors

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

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

Data as vectors

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

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

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

Data as vectors

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

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

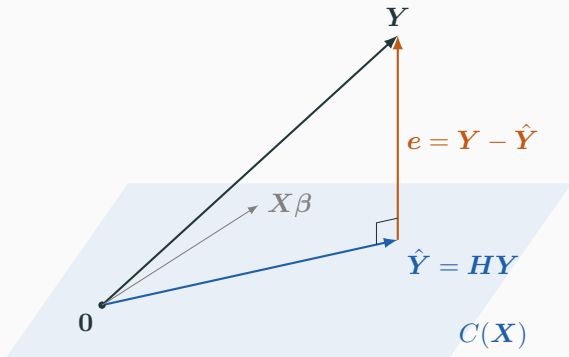
The model says that the mean $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ lies in $C(\mathbf{X})$. The observed $\mathbf{Y}$ is that mean plus noise, and is almost never in $C(\mathbf{X})$ itself.

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

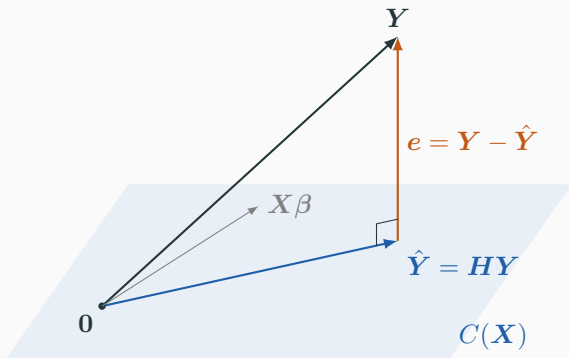
$$Q(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2,$$

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

Least squares is an orthogonal projection

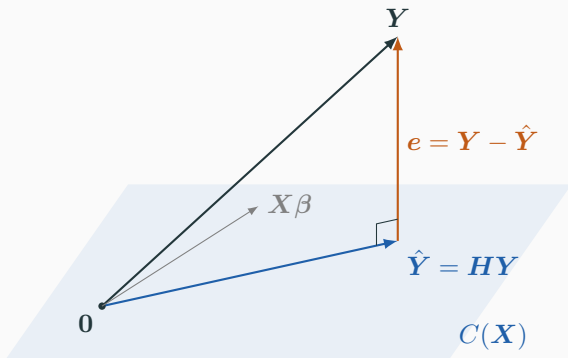


Least squares is an orthogonal projection



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

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.

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

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.

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

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

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.

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

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

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

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

$$\hat{Y} = HY.$$

The hat matrix is a projection

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

$$\hat{Y} = HY.$$

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

The hat matrix is a projection

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

$$\hat{Y} = HY.$$

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

The algebraic proofs are Assignment 2, Question 3.

The hat matrix is a projection

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

$$\hat{Y} = HY.$$

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

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 e belongs to this $(n - p)$ -dimensional space.

Properties of the residuals

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

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

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

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

Properties of the residuals

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

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

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

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

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

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

Properties of the residuals

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

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

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

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

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

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

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

Fitted values and residuals share the variance

Adding the two variance-covariance matrices:

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

Fitted values and residuals share the variance

Adding the two variance-covariance matrices:

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

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:

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

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:

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

For each observation:

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

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

$$Y_i = \beta_0 + \epsilon_i, \quad i = 1, \dots, n.$$

1. What is the design matrix $\mathbf{X}$?
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. $\mathbf{X} = \mathbf{1}$, the column of n ones, so $\mathbf{X}'\mathbf{X} = \mathbf{1}'\mathbf{1} = n$ and $(\mathbf{X}'\mathbf{X})^{-1} = 1/n$.

Exercise: solution

1. $\mathbf{X} = \mathbf{1}$, the column of n ones, so $\mathbf{X}'\mathbf{X} = \mathbf{1}'\mathbf{1} = n$ and $(\mathbf{X}'\mathbf{X})^{-1} = 1/n$.
2. $\mathbf{H} = \mathbf{1} \frac{1}{n} \mathbf{1}' = \frac{1}{n} \mathbf{J}$, where $\mathbf{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\mathbf{H}$ is $1/n$: $h_{ii} = 1/n$.

Exercise: solution

1. $\mathbf{X} = \mathbf{1}$, the column of n ones, so $\mathbf{X}'\mathbf{X} = \mathbf{1}'\mathbf{1} = n$ and $(\mathbf{X}'\mathbf{X})^{-1} = 1/n$.
2. $\mathbf{H} = \mathbf{1} \frac{1}{n} \mathbf{1}' = \frac{1}{n} \mathbf{J}$, where $\mathbf{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\mathbf{H}$ is $1/n$: $h_{ii} = 1/n$.
3. $\text{Var}(\hat{Y}_i) = h_{ii} \sigma^2 = \sigma^2/n$.

Exercise: solution

1. $\mathbf{X} = \mathbf{1}$, the column of n ones, so $\mathbf{X}'\mathbf{X} = \mathbf{1}'\mathbf{1} = n$ and $(\mathbf{X}'\mathbf{X})^{-1} = 1/n$.
2. $\mathbf{H} = \mathbf{1} \frac{1}{n} \mathbf{1}' = \frac{1}{n} \mathbf{J}$, where $\mathbf{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\mathbf{H}$ is $1/n$: $h_{ii} = 1/n$.
3. $\text{Var}(\hat{Y}_i) = h_{ii} \sigma^2 = \sigma^2/n$.
4. It is $\text{Var}(\bar{Y}) = \sigma^2/n$: indeed $\hat{\mathbf{Y}} = \frac{1}{n} \mathbf{J}\mathbf{Y} = \bar{Y}\mathbf{1}$, so $\hat{Y}_i = \bar{Y}$ for every i .

Exercise: solution

1. $\mathbf{X} = \mathbf{1}$, the column of n ones, so $\mathbf{X}'\mathbf{X} = \mathbf{1}'\mathbf{1} = n$ and $(\mathbf{X}'\mathbf{X})^{-1} = 1/n$.
2. $\mathbf{H} = \mathbf{1} \frac{1}{n} \mathbf{1}' = \frac{1}{n} \mathbf{J}$, where $\mathbf{J} = \mathbf{1}\mathbf{1}'$ is the $n \times n$ matrix of ones. Every entry of $\mathbf{H}$ is $1/n$: $h_{ii} = 1/n$.
3. $\text{Var}(\hat{Y}_i) = h_{ii} \sigma^2 = \sigma^2/n$.
4. It is $\text{Var}(\bar{Y}) = \sigma^2/n$: indeed $\hat{\mathbf{Y}} = \frac{1}{n} \mathbf{J}\mathbf{Y} = \bar{Y}\mathbf{1}$, 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.

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

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

Pythagoras and the analysis of variance

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

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

Pythagoras:

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

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

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

Pythagoras:

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

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
$\mathbf{Y} - \bar{Y}\mathbf{1}$	$n - 1$ (orthogonal to $\mathbf{1}$)	$n - 1$ for SST
$\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}$	$p - 1$ ($C(\mathbf{X})$, orthogonal to $\mathbf{1}$)	$p - 1$ for SSR
$\mathbf{e}$	$n - p$ (orthogonal to $C(\mathbf{X})$)	$n - p$ 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
$\mathbf{Y} - \bar{Y}\mathbf{1}$	$n - 1$ (orthogonal to $\mathbf{1}$)	$n - 1$ for SST
$\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}$	$p - 1$ ($C(\mathbf{X})$, orthogonal to $\mathbf{1}$)	$p - 1$ for SSR
$\mathbf{e}$	$n - p$ (orthogonal to $C(\mathbf{X})$)	$n - p$ for SSE

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

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

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

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
$\mathbf{Y} - \bar{Y}\mathbf{1}$	$n - 1$ (orthogonal to $\mathbf{1}$)	$n - 1$ for SST
$\hat{\mathbf{Y}} - \bar{Y}\mathbf{1}$	$p - 1$ ($C(\mathbf{X})$, orthogonal to $\mathbf{1}$)	$p - 1$ for SSR
$\mathbf{e}$	$n - p$ (orthogonal to $C(\mathbf{X})$)	$n - p$ for SSE

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

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

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

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

```
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

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

```
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

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.