STAC67: Regression Analysis

Lecture 15

Thibault Randrianarisoa

November 4, 2026

University of Toronto Scarborough

Review of Week 7

- Why the F test works. By the theorem on quadratic forms, $SSE/\sigma^2 \sim \chi^2_{n-p}$, independent of SSR/σ^2, a noncentral χ^2_{p-1} with $\Omega = \sum_i (\mu_i - \bar{\mu})^2/(2\sigma^2)$. Under H_0, $\Omega = 0$ and $MSR/MSE \sim F(p-1,\ n-p)$. Under H_a, F^* is pushed to the right.

- One linear combination $\theta = \boldsymbol{c}'\boldsymbol{\beta}$: $s^2\{\hat{\theta}\} = MSE\, \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$ and $(\hat{\theta} - \theta)/s\{\hat{\theta}\} \sim t(n-p)$.

- General linear hypothesis $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$, with q restrictions. Under H_0,

$$F^* = \frac{(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\left[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\right]^{-1}(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})\,/\,q}{MSE} \sim F(q,\ n-p).$$

  With one restriction ($q = 1$), $F^* = (t^*)^2$.

Today: extra sums of squares and multicollinearity, which we did not reach in Lecture 14, then qualitative predictors.

Extra sums of squares

The same test, from two fits

To test $H_0 : \beta_{\text{disp}} = \beta_{\text{cyl}} = 0$ in `acc ~ wt + hp + disp + cyl`, Lecture 14 took $\boldsymbol{K}' = [\,\boldsymbol{0} \mid \boldsymbol{I}_2\,]$ and $\boldsymbol{m} = \boldsymbol{0}$. Let $\hat{\boldsymbol{\beta}}_2$ be the $q = 2$ estimates tested, and $\boldsymbol{A} = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ their block of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

First form, from the full model (F) only:

$$F^* = \frac{\hat{\boldsymbol{\beta}}_2' \, \boldsymbol{A}^{-1} \hat{\boldsymbol{\beta}}_2 \,/\, q}{MSE} = 7.64.$$

The same test, from two fits

To test $H_0 : \beta_{\text{disp}} = \beta_{\text{cyl}} = 0$ in `acc ~ wt + hp + disp + cyl`, Lecture 14 took $\boldsymbol{K}' = [\,\mathbf{0} \mid \boldsymbol{I}_2\,]$ and $\boldsymbol{m} = \mathbf{0}$. Let $\hat{\boldsymbol{\beta}}_2$ be the $q = 2$ estimates tested, and $\boldsymbol{A} = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ their block of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

First form, from the full model (F) only:

$$F^* = \frac{\hat{\boldsymbol{\beta}}_2' \, \boldsymbol{A}^{-1} \hat{\boldsymbol{\beta}}_2 \,/\, q}{MSE} = 7.64.$$

Second form: also fit the reduced model (R) obtained by imposing H_0, `acc ~ wt + hp`. It is a constrained version of the full model, so it fits less well: $SSE(R) \geq SSE(F)$. Is the increase large?

$$F^* = \frac{\big(SSE(R) - SSE(F)\big) \,/\, q}{MSE} = \frac{(1174.3 - 1129.5)/2}{1129.5/385} = 7.64.$$

The same test, from two fits

To test $H_0 : \beta_{\text{disp}} = \beta_{\text{cyl}} = 0$ in `acc ~ wt + hp + disp + cyl`, Lecture 14 took $\boldsymbol{K}' = [\,\mathbf{0} \mid \boldsymbol{I}_2\,]$ and $\boldsymbol{m} = \mathbf{0}$. Let $\hat{\boldsymbol{\beta}}_2$ be the $q = 2$ estimates tested, and $\boldsymbol{A} = \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ their block of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

First form, from the full model (F) only:

$$F^* = \frac{\hat{\boldsymbol{\beta}}_2' \, \boldsymbol{A}^{-1} \hat{\boldsymbol{\beta}}_2 \,/\, q}{MSE} = 7.64.$$

Second form: also fit the reduced model (R) obtained by imposing H_0, `acc ~ wt + hp`. It is a constrained version of the full model, so it fits less well: $SSE(R) \geq SSE(F)$. Is the increase large?

$$F^* = \frac{\big(SSE(R) - SSE(F)\big) \,/\, q}{MSE} = \frac{(1174.3 - 1129.5)/2}{1129.5/385} = 7.64.$$

The denominator is the MSE of the full model in both, and q is the difference between the two error degrees of freedom. The numerators agree too: next slides. In R: `anova(reduced, full)`.

Why the numerators agree: Pythagoras

Split $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, with $\boldsymbol{X}_2$ the q predictors tested: the reduced model regresses $\boldsymbol{Y}$ on $\boldsymbol{X}_1$. Let

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1',$$

the $\boldsymbol{I} - \boldsymbol{H}$ of the reduced model: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$ and $SSE(R) = \|\boldsymbol{M}\boldsymbol{Y}\|^2$. With $\boldsymbol{e}$ the residuals of the full model, $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$, so $\boldsymbol{X}_2'\boldsymbol{e} = \boldsymbol{0}$ and $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}$.

Why the numerators agree: Pythagoras

Split $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, with $\boldsymbol{X}_2$ the q predictors tested: the reduced model regresses $\boldsymbol{Y}$ on $\boldsymbol{X}_1$. Let

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1',$$

the $\boldsymbol{I} - \boldsymbol{H}$ of the reduced model: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$ and $SSE(R) = \|\boldsymbol{M}\boldsymbol{Y}\|^2$. With $\boldsymbol{e}$ the residuals of the full model, $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$, so $\boldsymbol{X}_2'\boldsymbol{e} = \boldsymbol{0}$ and $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}$.

Apply $\boldsymbol{M}$ to the full fit $\boldsymbol{Y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}$:

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}.$$

Why the numerators agree: Pythagoras

Split $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$, with $\boldsymbol{X}_2$ the q predictors tested: the reduced model regresses $\boldsymbol{Y}$ on $\boldsymbol{X}_1$. Let

$$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1',$$

the $\boldsymbol{I} - \boldsymbol{H}$ of the reduced model: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$ and $SSE(R) = \|\boldsymbol{M}\boldsymbol{Y}\|^2$. With $\boldsymbol{e}$ the residuals of the full model, $\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0}$, so $\boldsymbol{X}_2'\boldsymbol{e} = \boldsymbol{0}$ and $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}$.

Apply $\boldsymbol{M}$ to the full fit $\boldsymbol{Y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}$:

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}.$$

The two terms are orthogonal, since $\hat{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{e} = \hat{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\boldsymbol{e} = 0$. By Pythagoras, $\|\boldsymbol{M}\boldsymbol{Y}\|^2 = \hat{\boldsymbol{\beta}}_2'\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \|\boldsymbol{e}\|^2$, that is

$$SSE(R) - SSE(F) = \hat{\boldsymbol{\beta}}_2'\left(\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2\right)\hat{\boldsymbol{\beta}}_2.$$

Why the numerators agree: the matrix A

It remains to show that $\boldsymbol{X}_2' \boldsymbol{M} \boldsymbol{X}_2 = \boldsymbol{A}^{-1}$.

Why the numerators agree: the matrix A

It remains to show that $\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2 = \boldsymbol{A}^{-1}$.

Multiply $\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}$ by $\boldsymbol{X}_2'$. Since $\boldsymbol{X}_2'\boldsymbol{e} = \boldsymbol{0}$,

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{Y}.$$

$\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2$ is invertible because $\boldsymbol{X}$ has full rank.

Why the numerators agree: the matrix A

It remains to show that $\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2 = \boldsymbol{A}^{-1}$.

Multiply $\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}$ by $\boldsymbol{X}_2'$. Since $\boldsymbol{X}_2'\boldsymbol{e} = \boldsymbol{0}$,

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{Y}.$$

$\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2$ is invertible because $\boldsymbol{X}$ has full rank.

With $\boldsymbol{\sigma^2}\{\boldsymbol{Y}\} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, this gives $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}_2\} = \sigma^2(\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2)^{-1}$. But $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}\} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ also gives $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}_2\} = \sigma^2\boldsymbol{A}$. Hence $\boldsymbol{A}^{-1} = \boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2$, and

$$SSE(R) - SSE(F) = \hat{\boldsymbol{\beta}}_2'\,\boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}_2.$$

Why the numerators agree: the matrix A

It remains to show that $\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2 = \boldsymbol{A}^{-1}$.

Multiply $\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{e}$ by $\boldsymbol{X}_2'$. Since $\boldsymbol{X}_2'\boldsymbol{e} = \boldsymbol{0}$,

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{Y}.$$

$\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2$ is invertible because $\boldsymbol{X}$ has full rank.

With $\boldsymbol{\sigma^2}\{\boldsymbol{Y}\} = \sigma^2\boldsymbol{I}$ and $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$, this gives $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}_2\} = \sigma^2(\boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2)^{-1}$. But $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}\} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ also gives $\boldsymbol{\sigma^2}\{\hat{\boldsymbol{\beta}}_2\} = \sigma^2\boldsymbol{A}$. Hence $\boldsymbol{A}^{-1} = \boldsymbol{X}_2'\boldsymbol{M}\boldsymbol{X}_2$, and

$$SSE(R) - SSE(F) = \hat{\boldsymbol{\beta}}_2'\,\boldsymbol{A}^{-1}\hat{\boldsymbol{\beta}}_2.$$

The two forms also agree for any $H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$: fitting the model under H_0 is another way to compute F^*.

Extra sums of squares

The drop in SSE when predictors are added is an extra sum of squares:

$$SSR(X_2 \mid X_1) = SSE(X_1) - SSE(X_1, X_2) = SSR(X_1, X_2) - SSR(X_1),$$

the variation explained by X_2 once X_1 is in the model. The two expressions agree because $SST = SSR + SSE$ for both models.

Extra sums of squares

The drop in SSE when predictors are added is an extra sum of squares:

$$SSR(X_2 \mid X_1) = SSE(X_1) - SSE(X_1, X_2) = SSR(X_1, X_2) - SSR(X_1),$$

the variation explained by X_2 once X_1 is in the model. The two expressions agree because $SST = SSR + SSE$ for both models.

With groups of predictors:
$SSR(X_3, X_4 \mid X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3, X_4)$, with 2 degrees of freedom.

Extra sums of squares

The drop in SSE when predictors are added is an extra sum of squares:

$$SSR(X_2 \mid X_1) = SSE(X_1) - SSE(X_1, X_2) = SSR(X_1, X_2) - SSR(X_1),$$

the variation explained by X_2 once X_1 is in the model. The two expressions agree because $SST = SSR + SSE$ for both models.

With groups of predictors:
$SSR(X_3, X_4 \mid X_1, X_2) = SSE(X_1, X_2) - SSE(X_1, X_2, X_3, X_4)$, with 2 degrees of freedom.

The test of $H_0 : \beta_3 = \beta_4 = 0$ is then

$$F^* = \frac{SSR(X_3, X_4 \mid X_1, X_2)/2}{MSE(X_1, X_2, X_3, X_4)} \sim F(2,\ n-5) \text{ under } H_0.$$

Decomposing SSR

Adding the predictors one at a time splits SSR into extra sums of squares, one degree of freedom each:

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2).$$

Decomposing SSR

Adding the predictors one at a time splits SSR into extra sums of squares, one degree of freedom each:

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2).$$

Source	SS	df
X_1	$SSR(X_1)$	1
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1
Error	$SSE(X_1, X_2, X_3)$	$n - 4$
Total	SST	$n - 1$

Decomposing SSR

Adding the predictors one at a time splits SSR into extra sums of squares, one degree of freedom each:

$$SSR(X_1, X_2, X_3) = SSR(X_1) + SSR(X_2 \mid X_1) + SSR(X_3 \mid X_1, X_2).$$

Source	SS	df
X_1	$SSR(X_1)$	1
$X_2 \mid X_1$	$SSR(X_2 \mid X_1)$	1
$X_3 \mid X_1, X_2$	$SSR(X_3 \mid X_1, X_2)$	1
Error	$SSE(X_1, X_2, X_3)$	$n - 4$
Total	SST	$n - 1$

⚠ This sequential decomposition depends on the order in which the predictors enter.

In R: `anova` gives the sequential sums of squares

```
anova(full)
anova(lm(acc ~ cyl + disp + hp + wt, data = autompg))
```

Predictor	Sequential SS, when the predictors enter in the order `wt, hp, disp, cyl`	`cyl, disp, hp, wt`
`wt`	522.3	334.7
`hp`	1272.4	615.2
`disp`	41.4	122.7
`cyl`	3.4	767.0
SSR (total of the four)	1839.5	1839.5
Error, SSE	1129.5	1129.5

In R: `anova` gives the sequential sums of squares

```
anova(full)
anova(lm(acc ~ cyl + disp + hp + wt, data = autompg))
```

Predictor	Sequential SS, when the predictors enter in the order `wt, hp, disp, cyl`	`cyl, disp, hp, wt`
`wt`	522.3	334.7
`hp`	1272.4	615.2
`disp`	41.4	122.7
`cyl`	3.4	767.0
SSR (total of the four)	1839.5	1839.5
Error, SSE	1129.5	1129.5

Weight explains 522.3 when it enters first, but only 334.7 once the other three are in: the predictors share much of the same information (their correlations are 0.84 to 0.95). SSR and SSE do not depend on the order.

Exercise: does the model year matter?

Add the model year to the acceleration model. From the two fits:

Model	SSE	error df
`acc ~ wt + hp`	1174.3	387
`acc ~ wt + hp + year`	1171.3	386

and in `summary` of the larger model, the t value of `year` is -0.990.

1. Test $H_0 : \beta_{\text{year}} = 0$ with the extra sum of squares, at the 5% level.
2. Check that your F^* is the square of the t value.
3. What does the conclusion say about the model year and acceleration?

Exercise: solution

1. $SSR(\texttt{year} \mid \texttt{wt}, \texttt{hp}) = 1174.3 - 1171.3 = 2.98$ on 1 df, and

$$F^* = \frac{2.98/1}{1171.3/386} = 0.98 < F(0.95;\ 1,\ 386) = 3.87 : \text{ do not reject } H_0.$$

Exercise: solution

1. $SSR(\texttt{year} \mid \texttt{wt}, \texttt{hp}) = 1174.3 - 1171.3 = 2.98$ on 1 df, and

$$F^* = \frac{2.98/1}{1171.3/386} = 0.98 < F(0.95;\ 1,\ 386) = 3.87 : \text{ do not reject } H_0.$$

2. $(-0.990)^2 = 0.98$: one restriction, so $F^* = (t^*)^2$, and the two tests are the same.

Exercise: solution

1. $SSR(\texttt{year} \mid \texttt{wt}, \texttt{hp}) = 1174.3 - 1171.3 = 2.98$ on 1 df, and

$$F^* = \frac{2.98/1}{1171.3/386} = 0.98 < F(0.95;\ 1,\ 386) = 3.87 : \text{ do not reject } H_0.$$

2. $(-0.990)^2 = 0.98$: one restriction, so $F^* = (t^*)^2$, and the two tests are the same.
3. Once weight and horsepower are in the model, the data show no evidence that the model year adds anything. This does not say that year is unrelated to acceleration on its own: newer cars were also lighter and less powerful.

Multicollinearity

When the predictors are uncorrelated

Eight work crews: crew size X_1, bonus pay X_2, productivity Y. The predictors are uncorrelated below: $r(X_1, X_2) = 0$.

X_1	4	4	4	4	6	6	6	6
X_2	2	2	3	3	2	2	3	3
Y	42	39	48	51	49	53	61	60

When the predictors are uncorrelated

Eight work crews: crew size X_1, bonus pay X_2, productivity Y. The predictors are uncorrelated below: $r(X_1, X_2) = 0$.

X_1	4	4	4	4	6	6	6	6
X_2	2	2	3	3	2	2	3	3
Y	42	39	48	51	49	53	61	60

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	5.375	
$Y \sim X_2$		9.250
$Y \sim X_1 + X_2$	5.375	9.250

When the predictors are uncorrelated

Eight work crews: crew size X_1, bonus pay X_2, productivity Y. The predictors are uncorrelated below: $r(X_1, X_2) = 0$.

X_1	4	4	4	4	6	6	6	6
X_2	2	2	3	3	2	2	3	3
Y	42	39	48	51	49	53	61	60

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	5.375	
$Y \sim X_2$		9.250
$Y \sim X_1 + X_2$	5.375	9.250

The slopes do not change, and neither do the extra sums of squares: $SSR(X_1) = SSR(X_1 \mid X_2) = 231.125$ and $SSR(X_2) = SSR(X_2 \mid X_1) = 171.125$. The order no longer matters.

When the predictors are correlated

In the acceleration data, the four predictors move together:

```
       wt   hp disp  cyl
wt   1.00 0.86 0.93 0.90
hp   0.86 1.00 0.90 0.84
disp 0.93 0.90 1.00 0.95
cyl  0.90 0.84 0.95 1.00
```

When the predictors are correlated

In the acceleration data, the four predictors move together:

```
       wt   hp disp  cyl
wt   1.00 0.86 0.93 0.90
hp   0.86 1.00 0.90 0.84
disp 0.93 0.90 1.00 0.95
cyl  0.90 0.84 0.95 1.00
```

The coefficient of weight depends on which other predictors are in the model:

Model	$\hat{\beta}_{\text{wt}}$	$s\{\hat{\beta}_{\text{wt}}\}$
`acc ~ wt`	-0.00136	0.00015
`acc ~ wt + hp`	0.00229	0.00021
`acc ~ wt + hp + disp`	0.00306	0.00029
`acc ~ wt + hp + disp + cyl`	0.00309	0.00029

When the predictors are correlated

In the acceleration data, the four predictors move together:

```
       wt   hp disp  cyl
wt   1.00 0.86 0.93 0.90
hp   0.86 1.00 0.90 0.84
disp 0.93 0.90 1.00 0.95
cyl  0.90 0.84 0.95 1.00
```

The coefficient of weight depends on which other predictors are in the model:

Model	$\hat{\beta}_{\text{wt}}$	$s\{\hat{\beta}_{\text{wt}}\}$
`acc ~ wt`	-0.00136	0.00015
`acc ~ wt + hp`	0.00229	0.00021
`acc ~ wt + hp + disp`	0.00306	0.00029
`acc ~ wt + hp + disp + cyl`	0.00309	0.00029

The sign changes once horsepower is in (Lecture 10), and the standard error doubles as more correlated predictors are added.

Why the variances inflate

With two predictors, the variance of a slope is (stated without proof)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}\,(1 - r_{12}^2)}, \qquad S_{11} = \sum_i (X_{i1} - \bar{X}_1)^2,$$

where r_{12} is the correlation between the two predictors. With one predictor, it was σ^2 / S_{11}.

Why the variances inflate

With two predictors, the variance of a slope is (stated without proof)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}\,(1 - r_{12}^2)}, \qquad S_{11} = \sum_i (X_{i1} - \bar{X}_1)^2,$$

where r_{12} is the correlation between the two predictors. With one predictor, it was σ^2 / S_{11}.

The factor $1/(1 - r_{12}^2)$ is 1 when $r_{12} = 0$, and explodes as $|r_{12}| \to 1$.

Displacement and cylinders: $r = 0.951$, a factor of 10.4 on the variance:

	`acc ~ disp`	`acc ~ disp + cyl`
$s\{\hat{\beta}_{\text{disp}}\}$	0.0011	0.0036

The standard error is multiplied by about $\sqrt{10.4} \approx 3.2$.

Why the variances inflate

With two predictors, the variance of a slope is (stated without proof)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}\,(1 - r_{12}^2)}, \qquad S_{11} = \sum_i (X_{i1} - \bar{X}_1)^2,$$

where r_{12} is the correlation between the two predictors. With one predictor, it was σ^2 / S_{11}.

The factor $1/(1 - r_{12}^2)$ is 1 when $r_{12} = 0$, and explodes as $|r_{12}| \to 1$.

Displacement and cylinders: $r = 0.951$, a factor of 10.4 on the variance:

	`acc ~ disp`	`acc ~ disp + cyl`
$s\{\hat{\beta}_{\text{disp}}\}$	0.0011	0.0036

The standard error is multiplied by about $\sqrt{10.4} \approx 3.2$.

With more predictors, $1/(1 - R_k^2)$, where R_k^2 comes from regressing X_k on the other predictors, is the *variance inflation factor*, later in the course. For displacement among the four predictors, it is 19.3.

Significant alone, but not once the others are in

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.0553	0.4061	49.3866	0
cyl	-0.8226	0.0708	-11.6248	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.8864	0.6018	29.7228	0.0000
disp	-0.0173	0.0036	-4.7785	0.0000
cyl	0.1884	0.2225	0.8467	0.3977

Significant alone, but not once the others are in

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.0553	0.4061	49.3866	0
cyl	-0.8226	0.0708	-11.6248	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.8864	0.6018	29.7228	0.0000
disp	-0.0173	0.0036	-4.7785	0.0000
cyl	0.1884	0.2225	0.8467	0.3977

Alone, cylinders are strongly related to acceleration ($t = -11.6$). Once displacement is in the model, their coefficient is not significant ($t = 0.85$, $p = 0.40$), and its sign even flips.

Significant alone, but not once the others are in

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.0553	0.4061	49.3866	0
cyl	-0.8226	0.0708	-11.6248	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.8864	0.6018	29.7228	0.0000
disp	-0.0173	0.0036	-4.7785	0.0000
cyl	0.1884	0.2225	0.8467	0.3977

Alone, cylinders are strongly related to acceleration ($t = -11.6$). Once displacement is in the model, their coefficient is not significant ($t = 0.85$, $p = 0.40$), and its sign even flips.

The t test asks whether cylinders add anything once displacement is known. With $r = 0.95$, they carry almost the same information, so the answer is: very little.

What multicollinearity does, and does not do

It does not violate any assumption: $\hat{\beta}$ is still unbiased, and, unless the dependence is exact, $\boldsymbol{X}'\boldsymbol{X}$ can still be inverted. Fitted values and predictions inside the range of the data stay valid.

What multicollinearity does, and does not do

It does not violate any assumption: $\hat{\beta}$ is still unbiased, and, unless the dependence is exact, $\boldsymbol{X}'\boldsymbol{X}$ can still be inverted. Fitted values and predictions inside the range of the data stay valid.

It does:

- inflate the standard errors of the coefficients;
- make the coefficients unstable: they change, and can change sign, depending on which predictors are in the model;
- blur the meaning of "the others held fixed", when the predictors rarely move separately;
- make the sequential sums of squares depend heavily on the order;
- allow every individual t test to be non-significant while the overall F test rejects.

What multicollinearity does, and does not do

It does not violate any assumption: $\hat{\beta}$ is still unbiased, and, unless the dependence is exact, $\boldsymbol{X}'\boldsymbol{X}$ can still be inverted. Fitted values and predictions inside the range of the data stay valid.

It does:

- inflate the standard errors of the coefficients;
- make the coefficients unstable: they change, and can change sign, depending on which predictors are in the model;
- blur the meaning of "the others held fixed", when the predictors rarely move separately;
- make the sequential sums of squares depend heavily on the order;
- allow every individual t test to be non-significant while the overall F test rejects.

Remedies, later in the course: drop or combine predictors, or use ridge regression.

Qualitative predictors

Coding a qualitative predictor

Many predictors are qualitative (categorical): the type of a company, a treatment group, a region. They enter the model through indicator variables (dummy variables):

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ is in the group,} \\ 0 & \text{otherwise.} \end{cases}$$

Coding a qualitative predictor

Many predictors are qualitative (categorical): the type of a company, a treatment group, a region. They enter the model through indicator variables (dummy variables):

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ is in the group,} \\ 0 & \text{otherwise.} \end{cases}$$

With a single binary predictor, the model is

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i : \qquad E(Y_i) = \begin{cases} \beta_0 & \text{if } D_i = 0, \\ \beta_0 + \beta_1 & \text{if } D_i = 1. \end{cases}$$

β_0 is the mean response of the group coded 0, and β_1 is the difference between the two group means.

Coding a qualitative predictor

Many predictors are qualitative (categorical): the type of a company, a treatment group, a region. They enter the model through indicator variables (dummy variables):

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ is in the group,} \\ 0 & \text{otherwise.} \end{cases}$$

With a single binary predictor, the model is

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i : \qquad E(Y_i) = \begin{cases} \beta_0 & \text{if } D_i = 0, \\ \beta_0 + \beta_1 & \text{if } D_i = 1. \end{cases}$$

β_0 is the mean response of the group coded 0, and β_1 is the difference between the two group means.

Least squares gives $\hat{\beta}_0 = \bar{Y}_0$ and $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$, the sample means of the groups. The t test of $H_0 : \beta_1 = 0$ is the two-sample t test with a pooled variance.

Example: how fast do insurance firms innovate?

For 20 insurance firms: the number of months before the firm adopted an innovation (`months`), and whether it is a stock company ($D = 1$, 10 firms) or a mutual company ($D = 0$, 10 firms).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.7	2.92	5.719	0.000
stock	5.4	4.13	1.308	0.207

Example: how fast do insurance firms innovate?

For 20 insurance firms: the number of months before the firm adopted an innovation (`months`), and whether it is a stock company ($D = 1$, 10 firms) or a mutual company ($D = 0$, 10 firms).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.7	2.92	5.719	0.000
stock	5.4	4.13	1.308	0.207

$\hat{\beta}_0 = 16.7$ months is the mean of the mutual firms, and $\hat{\beta}_1 = 5.4$ months is the difference: stock firms took longer, on average.

Example: how fast do insurance firms innovate?

For 20 insurance firms: the number of months before the firm adopted an innovation (`months`), and whether it is a stock company ($D = 1$, 10 firms) or a mutual company ($D = 0$, 10 firms).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.7	2.92	5.719	0.000
stock	5.4	4.13	1.308	0.207

$\hat{\beta}_0 = 16.7$ months is the mean of the mutual firms, and $\hat{\beta}_1 = 5.4$ months is the difference: stock firms took longer, on average.

But $t = 1.31$ ($p = 0.21$): on its own, the type of firm shows no significant effect. We come back to this after adding the size of the firm.

A factor with k levels

A qualitative predictor with k levels needs $k-1$ indicators. The level without an indicator is the baseline: its mean is β_0, and each coefficient is the difference between a level and the baseline:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \cdots + \beta_{k-1} D_{i,k-1} + \epsilon_i,$$

$$E(Y_i) = \begin{cases} \beta_0 & \text{at the baseline,} \\ \beta_0 + \beta_j & \text{at the level of } D_j. \end{cases}$$

A factor with k levels

A qualitative predictor with k levels needs $k-1$ indicators. The level without an indicator is the baseline: its mean is β_0, and each coefficient is the difference between a level and the baseline:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \cdots + \beta_{k-1} D_{i,k-1} + \epsilon_i,$$

$$E(Y_i) = \begin{cases} \beta_0 & \text{at the baseline,} \\ \beta_0 + \beta_j & \text{at the level of } D_j. \end{cases}$$

Why not k indicators? With an intercept, the k indicators add up to the column of 1s: $D_{i1} + \cdots + D_{ik} = 1$ for every i. The columns of $\boldsymbol{X}$ are linearly dependent and $\boldsymbol{X}'\boldsymbol{X}$ cannot be inverted (Problem Set 5, Exercise 3).

A factor with k levels

A qualitative predictor with k levels needs $k-1$ indicators. The level without an indicator is the baseline: its mean is β_0, and each coefficient is the difference between a level and the baseline:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \cdots + \beta_{k-1} D_{i,k-1} + \epsilon_i,$$

$$E(Y_i) = \begin{cases} \beta_0 & \text{at the baseline,} \\ \beta_0 + \beta_j & \text{at the level of } D_j. \end{cases}$$

Why not k indicators? With an intercept, the k indicators add up to the column of 1s: $D_{i1} + \cdots + D_{ik} = 1$ for every i. The columns of $\boldsymbol{X}$ are linearly dependent and $\boldsymbol{X}'\boldsymbol{X}$ cannot be inverted (Problem Set 5, Exercise 3).

The choice of baseline does not change the fitted values; it only changes which differences the coefficients report. R takes the first level in alphabetical order.

Example: plant weights in three groups

Dried weights of 30 plants: a control group and two treatments, 10 plants each (`PlantGrowth`, in R).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.032	0.197	25.527	0.000
grouptrt1	-0.371	0.279	-1.331	0.194
grouptrt2	0.494	0.279	1.772	0.088

Example: plant weights in three groups

Dried weights of 30 plants: a control group and two treatments, 10 plants each (`PlantGrowth`, in R).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.032	0.197	25.527	0.000
grouptrt1	-0.371	0.279	-1.331	0.194
grouptrt2	0.494	0.279	1.772	0.088

Baseline: `ctrl`, with mean 5.032. The coefficients are the differences `trt1` $-$ `ctrl` $= -0.371$ and `trt2` $-$ `ctrl` $= 0.494$.

Example: plant weights in three groups

Dried weights of 30 plants: a control group and two treatments, 10 plants each (`PlantGrowth`, in R).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.032	0.197	25.527	0.000
grouptrt1	-0.371	0.279	-1.331	0.194
grouptrt2	0.494	0.279	1.772	0.088

Baseline: `ctrl`, with mean 5.032. The coefficients are the differences `trt1` $-$ `ctrl` $= -0.371$ and `trt2` $-$ `ctrl` $= 0.494$.

Are the three group means equal? $H_0 : \beta_1 = \beta_2 = 0$ is the overall F test of this model: $F^* = 4.85$ on 2 and 27 df, $p = 0.016$. We reject: the group means are not all equal (here `trt1` and `trt2` lie on either side of the control).

This is the one-way analysis of variance.

Why not code the levels $1, 2, \ldots, k$?

Back to the plants: code the groups `ctrl`, `trt1` and `trt2` as $L = 1, 2, 3$ in one column:

$$Y_i = \beta_0 + \beta_1 L_i + \epsilon_i : \qquad E(Y_i) = \beta_0 + \beta_1, \quad \beta_0 + 2\beta_1, \quad \beta_0 + 3\beta_1.$$

Going from one group to the next always adds the same amount β_1: the levels are forced to be ordered and equally spaced.

Why not code the levels $1, 2, \ldots, k$?

Back to the plants: code the groups `ctrl`, `trt1` and `trt2` as $L = 1, 2, 3$ in one column:

$$Y_i = \beta_0 + \beta_1 L_i + \epsilon_i : \qquad E(Y_i) = \beta_0 + \beta_1, \quad \beta_0 + 2\beta_1, \quad \beta_0 + 3\beta_1.$$

Going from one group to the next always adds the same amount β_1: the levels are forced to be ordered and equally spaced.

For these data:

	`ctrl`	`trt1`	`trt2`
Sample means	5.03	4.66	5.53
Fitted means, codes $1, 2, 3$	4.83	5.07	5.32
Fitted means, codes $2, 1, 3$	5.07	4.64	5.51

Why not code the levels $1, 2, \ldots, k$?

Back to the plants: code the groups `ctrl`, `trt1` and `trt2` as $L = 1, 2, 3$ in one column:

$$Y_i = \beta_0 + \beta_1 L_i + \epsilon_i : \qquad E(Y_i) = \beta_0 + \beta_1, \quad \beta_0 + 2\beta_1, \quad \beta_0 + 3\beta_1.$$

Going from one group to the next always adds the same amount β_1: the levels are forced to be ordered and equally spaced.

For these data:

	`ctrl`	`trt1`	`trt2`
Sample means	5.03	4.66	5.53
Fitted means, codes $1, 2, 3$	4.83	5.07	5.32
Fitted means, codes $2, 1, 3$	5.07	4.64	5.51

The fit misses the means, and it changes when the labels, which are arbitrary, are swapped. With indicators, $E(Y_i) = \beta_0,\ \beta_0 + \beta_1,\ \beta_0 + \beta_2$: each level has its own mean, and the fitted means are the sample means, whatever the labels.

Summary

- Extra sums of squares measure what predictors add once others are in, and give the F test from a full and a reduced model. The sequential decomposition depends on the order, unless the predictors are uncorrelated.
- Correlated predictors inflate the variances of the coefficients and make them hard to interpret, without violating any assumption.
- A qualitative predictor with k levels enters through $k - 1$ indicators in the model. Each coefficient compares a level with the baseline group.
- The model estimates the group means, and testing if theyr are all equal is the one-way analysis of variance.

Next: a qualitative predictor next to a quantitative one.