STAC67: Regression Analysis

Lecture 14

Thibault Randrianarisoa

October 23, 2026

University of Toronto Scarborough

Review of Wednesday

The overall F test of $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$ uses $F^* = MSR/MSE \sim F(p-1,\ n-p)$ under H_0, and

$$F^* = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}, \qquad R^2 = \frac{SSR}{SST}, \qquad R_a^2 = 1 - \frac{n-1}{n-p}\frac{SSE}{SST}.$$

Noncentral χ^2: if $Z_1, \ldots, Z_r$ are independent with $Z_i \sim N(\mu_i, 1)$, then $W = \sum_{i=1}^r Z_i^2$ has a noncentral χ^2 distribution with r degrees of freedom and noncentrality parameter $\Omega = \frac{1}{2}\sum_i \mu_i^2$.

- $\Omega = 0$: the usual χ^2_r. $E(W) = r + 2\Omega$: the larger Ω, the further to the right.

Why the F test works

The distribution of quadratic forms

Until now, the F distribution of MSR/MSE has been takem for granted.

Theorem

Let $\boldsymbol{Z} \sim N(\boldsymbol{\mu},\, \sigma^2 \boldsymbol{V})$, with $\boldsymbol{V}$ nonsingular, and let $\boldsymbol{A}$ and $\boldsymbol{B}$ be constant matrices, $\boldsymbol{A}$ symmetric.

1. If $\boldsymbol{AV}$ is idempotent, then $\boldsymbol{Z}'\boldsymbol{AZ}/\sigma^2$ has a noncentral χ^2 distribution (recalled in the review), with $r(\boldsymbol{A})$ degrees of freedom (the rank of $\boldsymbol{A}$) and noncentrality parameter $\Omega = \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}/(2\sigma^2)$. If $\Omega = 0$, it is the usual $\chi^2_{r(\boldsymbol{A})}$.
2. $\boldsymbol{Z}'\boldsymbol{AZ}$ and $\boldsymbol{Z}'\boldsymbol{BZ}$ ($\boldsymbol{B}$ symmetric) are independent if $\boldsymbol{AVB} = \boldsymbol{0}$.
3. $\boldsymbol{Z}'\boldsymbol{AZ}$ and $\boldsymbol{BZ}$ are independent if $\boldsymbol{BVA} = \boldsymbol{0}$.

For the response variable, $\boldsymbol{Z} = \boldsymbol{Y}$, $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{V} = \boldsymbol{I}$: the conditions become $\boldsymbol{AA} = \boldsymbol{A}$ and $\boldsymbol{AB} = \boldsymbol{0}$.

Applying it to SSE and SSR

$\boldsymbol{Y} \sim N(\boldsymbol{\mu},\, \sigma^2 \boldsymbol{I})$ with $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, $SSE = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $SSR = \boldsymbol{Y}'(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ (Lecture 12).

Applying it to SSE and SSR

$\boldsymbol{Y} \sim N(\boldsymbol{\mu},\, \sigma^2 \boldsymbol{I})$ with $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, $SSE = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $SSR = \boldsymbol{Y}'(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ (Lecture 12).

Both matrices are idempotent (Lecture 12), their ranks are their traces, $n - p$ and $p - 1$, and their product is $\mathbf{0}$, since $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ and $\boldsymbol{J}\boldsymbol{H} = \boldsymbol{J}$:

$$\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{H} - \boldsymbol{H}\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J} + \tfrac{1}{n}\boldsymbol{J}\boldsymbol{H} = \mathbf{0}.$$

Applying it to SSE and SSR

$\boldsymbol{Y} \sim N(\boldsymbol{\mu},\, \sigma^2 \boldsymbol{I})$ with $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, $SSE = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ and $SSR = \boldsymbol{Y}'(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y}$ (Lecture 12).

Both matrices are idempotent (Lecture 12), their ranks are their traces, $n - p$ and $p - 1$, and their product is $\mathbf{0}$, since $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ and $\boldsymbol{J}\boldsymbol{H} = \boldsymbol{J}$:

$$\left(\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J}\right)(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{H} - \boldsymbol{H}\boldsymbol{H} - \tfrac{1}{n}\boldsymbol{J} + \tfrac{1}{n}\boldsymbol{J}\boldsymbol{H} = \mathbf{0}.$$

- $SSE/\sigma^2 \sim \chi^2_{n-p}$: its Ω is 0, because $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \mathbf{0}$.
- SSR/σ^2 is noncentral χ^2 with $p - 1$ degrees of freedom and $\Omega = \sum_i (\mu_i - \bar{\mu})^2/(2\sigma^2)$, since $\boldsymbol{H}\boldsymbol{\mu} = \boldsymbol{\mu}$ gives $\boldsymbol{\mu}'(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{\mu} - n\bar{\mu}^2 = \sum_i (\mu_i - \bar{\mu})^2$.
- SSR and SSE are independent (part 2 of the theorem).

The overall F test, justified

Under $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$, every $\mu_i = \beta_0$, so $\Omega = 0$ and $SSR/\sigma^2 \sim \chi^2_{p-1}$, independent of $SSE/\sigma^2 \sim \chi^2_{n-p}$. Hence

$$F^* = \frac{MSR}{MSE} = \frac{\big(SSR/\sigma^2\big)/(p-1)}{\big(SSE/\sigma^2\big)/(n-p)} \sim F(p-1,\, n-p) \quad \text{under } H_0 :$$

a ratio of two independent χ^2 variables, each divided by its degrees of freedom. σ^2 cancels, so F^* can be computed.

The overall F test, justified

Under $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$, every $\mu_i = \beta_0$, so $\Omega = 0$ and $SSR/\sigma^2 \sim \chi^2_{p-1}$, independent of $SSE/\sigma^2 \sim \chi^2_{n-p}$. Hence

$$F^* = \frac{MSR}{MSE} = \frac{\big(SSR/\sigma^2\big)/(p-1)}{\big(SSE/\sigma^2\big)/(n-p)} \sim F(p-1,\, n-p) \quad \text{under } H_0 :$$

a ratio of two independent χ^2 variables, each divided by its degrees of freedom. σ^2 cancels, so F^* can be computed.

Under H_a, the μ_i differ and $\Omega > 0$: SSR/σ^2 is pushed to the right, and F^* with it. This is why the test rejects only for large values of F^*, as $E(MSR) \geq \sigma^2$ suggested last Friday.

One linear combination of the coefficients

Estimating $\boldsymbol{c}'\boldsymbol{\beta}$

Many questions are about a linear combinations of the parameters

$$\theta = \boldsymbol{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \cdots + c_{p-1}\beta_{p-1},$$

for some constant vector $\boldsymbol{c}$:

- $\boldsymbol{c} = (0, \ldots, 1, \ldots, 0)'$: a single coefficient β_k;
- $\boldsymbol{c} = \boldsymbol{X}_h$: the mean response at $\boldsymbol{X}_h$ (Lecture 11);
- $\boldsymbol{c} = \boldsymbol{X}_{h_1} - \boldsymbol{X}_{h_2}$: the difference between two mean responses.

Estimating $c'\beta$

Many questions are about a linear combinations of the parameters

$$\theta = \boldsymbol{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \cdots + c_{p-1}\beta_{p-1},$$

for some constant vector $\boldsymbol{c}$:

- $\boldsymbol{c} = (0, \ldots, 1, \ldots, 0)'$: a single coefficient β_k;
- $\boldsymbol{c} = \boldsymbol{X}_h$: the mean response at $\boldsymbol{X}_h$ (Lecture 11);
- $\boldsymbol{c} = \boldsymbol{X}_{h_1} - \boldsymbol{X}_{h_2}$: the difference between two mean responses.

It is estimated by $\hat{\theta} = \boldsymbol{c}'\hat{\boldsymbol{\beta}}$, a linear function of $\hat{\boldsymbol{\beta}}$:

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

Estimating $c'\beta$

Many questions are about a linear combinations of the parameters

$$\theta = \boldsymbol{c}'\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + \cdots + c_{p-1}\beta_{p-1},$$

for some constant vector $\boldsymbol{c}$:

- $\boldsymbol{c} = (0, \ldots, 1, \ldots, 0)'$: a single coefficient β_k;
- $\boldsymbol{c} = \boldsymbol{X}_h$: the mean response at $\boldsymbol{X}_h$ (Lecture 11);
- $\boldsymbol{c} = \boldsymbol{X}_{h_1} - \boldsymbol{X}_{h_2}$: the difference between two mean responses.

It is estimated by $\hat{\theta} = \boldsymbol{c}'\hat{\boldsymbol{\beta}}$, a linear function of $\hat{\boldsymbol{\beta}}$:

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

With normal errors, and $s^2\{\hat{\theta}\} = MSE\,\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$: $\dfrac{\hat{\theta} - \theta}{s\{\hat{\theta}\}} \sim t(n-p)$.

Hence the interval $\hat{\theta} \pm t(1-\alpha/2;\, n-p)\, s\{\hat{\theta}\}$, and the test of $H_0 : \theta = \theta_0$ with $t^* = (\hat{\theta} - \theta_0)/s\{\hat{\theta}\}$ follow as before.

Example: two cars

Model $\texttt{acc} = \beta_0 + \beta_1\,\texttt{wt} + \beta_2\,\texttt{hp} + \epsilon$. How much faster, on average, is a car of 3500 lb and 130 hp than one of 3000 lb and 100 hp?

$$\theta = E(Y \mid 3500, 130) - E(Y \mid 3000, 100) = 500\,\beta_1 + 30\,\beta_2, \qquad \boldsymbol{c} = (0, 500, 30)'.$$

Recall that `vcov(fit2)` is $MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

```
fit2 <- lm(acc ~ wt + hp, data = autompg)
cc <- c(0, 500, 30)
theta_hat <- sum(cc * coef(fit2))
s_theta <- sqrt(drop(t(cc) %*% vcov(fit2) %*% cc))
c(theta_hat = theta_hat, s_theta = s_theta)
```

```
##   theta_hat     s_theta
## -1.65182813  0.07030124
```

Example: two cars

Model $\texttt{acc} = \beta_0 + \beta_1\,\texttt{wt} + \beta_2\,\texttt{hp} + \epsilon$. How much faster, on average, is a car of 3500 lb and 130 hp than one of 3000 lb and 100 hp?

$$\theta = E(Y \mid 3500, 130) - E(Y \mid 3000, 100) = 500\,\beta_1 + 30\,\beta_2, \qquad \boldsymbol{c} = (0, 500, 30)'.$$

Recall that `vcov(fit2)` is $MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

```
fit2 <- lm(acc ~ wt + hp, data = autompg)
cc <- c(0, 500, 30)
theta_hat <- sum(cc * coef(fit2))
s_theta <- sqrt(drop(t(cc) %*% vcov(fit2) %*% cc))
c(theta_hat = theta_hat, s_theta = s_theta)
```

```
##   theta_hat     s_theta
## -1.65182813  0.07030124
```

95% interval: $-1.65 \pm 1.966 \times 0.070 = (-1.79,\ -1.51)$ s. The heavier but more powerful car reaches 60 mph 1.5 to 1.8 s sooner on average.

The general linear hypothesis

Several restrictions at once

A general linear hypothesis imposes q linear restrictions on $\boldsymbol{\beta}$:

$$H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m} \qquad \text{against} \qquad H_a : \boldsymbol{K}'\boldsymbol{\beta} \neq \boldsymbol{m},$$

where $\boldsymbol{K}'$ is a constant $q \times p$ matrix of rank q (one row per restriction) and $\boldsymbol{m}$ is a constant $q \times 1$ vector.

Several restrictions at once

A general linear hypothesis imposes q linear restrictions on $\boldsymbol{\beta}$:

$$H_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m} \qquad \text{against} \qquad H_a : \boldsymbol{K}'\boldsymbol{\beta} \neq \boldsymbol{m},$$

where $\boldsymbol{K}'$ is a constant $q \times p$ matrix of rank q (one row per restriction) and $\boldsymbol{m}$ is a constant $q \times 1$ vector.

Examples, with $p = 5$ parameters $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$:

- $H_0 : \beta_3 = \beta_4 = 0$: $\boldsymbol{K}' = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{m} = \boldsymbol{0},\ q = 2.$
- $H_0 : \beta_1 = \beta_2$: $\boldsymbol{K}' = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix}, \boldsymbol{m} = 0,\ q = 1.$
- The overall test: $\boldsymbol{K}' = [\,\boldsymbol{0} \mid \boldsymbol{I}_4\,], \boldsymbol{m} = \boldsymbol{0},\ q = p - 1 = 4.$

The F statistic

$\boldsymbol{K}'\hat{\boldsymbol{\beta}}$ is a linear function of $\hat{\boldsymbol{\beta}}$. Under H_0, with normal errors,

$$\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m} \sim N\big(\mathbf{0},\ \sigma^2\, \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big).$$

The F statistic

$\boldsymbol{K}'\hat{\boldsymbol{\beta}}$ is a linear function of $\hat{\boldsymbol{\beta}}$. Under H_0, with normal errors,

$$\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m} \sim N\big(\mathbf{0},\ \sigma^2\, \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big).$$

The theorem on quadratic forms, with $\boldsymbol{A}$ the inverse of $\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ (so $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{I}$), gives a χ^2_q variable below, independent of SSE since $\hat{\boldsymbol{\beta}}$ is (part 3: $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}) = \mathbf{0}$).

Meaning

$$F^* = \frac{(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\big[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big]^{-1}(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})\,/\,q}{MSE} \sim F(q,\, n{-}p) \text{ under } H_0.$$

The F statistic

$\boldsymbol{K}'\hat{\boldsymbol{\beta}}$ is a linear function of $\hat{\boldsymbol{\beta}}$. Under H_0, with normal errors,

$$\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m} \sim N\big(\mathbf{0},\ \sigma^2\, \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big).$$

The theorem on quadratic forms, with $\boldsymbol{A}$ the inverse of $\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ (so $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{I}$), gives a χ^2_q variable below, independent of SSE since $\hat{\boldsymbol{\beta}}$ is (part 3: $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}) = \mathbf{0}$).

Meaning

$$F^* = \frac{(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\big[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big]^{-1}(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})\,/\,q}{MSE} \sim F(q,\, n{-}p) \text{ under } H_0.$$

Reject H_0 at level α when $F^* > F(1-\alpha;\ q,\ n-p)$.

The F statistic

$\boldsymbol{K}'\hat{\boldsymbol{\beta}}$ is a linear function of $\hat{\boldsymbol{\beta}}$. Under H_0, with normal errors,

$$\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m} \sim N\big(\mathbf{0},\ \sigma^2\, \boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big).$$

The theorem on quadratic forms, with $\boldsymbol{A}$ the inverse of $\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}$ (so $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{I}$), gives a χ^2_q variable below, independent of SSE since $\hat{\boldsymbol{\beta}}$ is (part 3: $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{H}) = \mathbf{0}$).

Meaning

$$F^* = \frac{(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})'\big[\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K}\big]^{-1}(\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m})\,/\,q}{MSE} \sim F(q,\, n{-}p) \text{ under } H_0.$$

Reject H_0 at level α when $F^* > F(1-\alpha;\ q,\ n-p)$.

With one restriction ($q = 1$), F^* is the square of the t^*, the student statistic.

In R: do displacement and cylinders matter?

Full model: `acc ~ wt + hp + disp + cyl`.

$H_0 : \beta_{\text{disp}} = \beta_{\text{cyl}} = 0$, so $q = 2$.

Since `vcov(full)` is $MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$, the numerator and MSE combine:

```
full <- lm(acc ~ wt + hp + disp + cyl, data = autompg)
Kt <- rbind(c(0, 0, 0, 1, 0),
            c(0, 0, 0, 0, 1))
d <- Kt %*% coef(full)            # here m = 0
V <- Kt %*% vcov(full) %*% t(Kt)
F_star <- drop(t(d) %*% solve(V) %*% d) / 2
c(F_star, 1 - pf(F_star, 2, 385))
```

```
## [1] 7.6362147397 0.0005594001
```

In R: do displacement and cylinders matter?

Full model: `acc ~ wt + hp + disp + cyl`.

$H_0 : \beta_{\text{disp}} = \beta_{\text{cyl}} = 0$, so $q = 2$.

Since `vcov(full)` is $MSE\,(\boldsymbol{X}'\boldsymbol{X})^{-1}$, the numerator and MSE combine:

```
full <- lm(acc ~ wt + hp + disp + cyl, data = autompg)
Kt <- rbind(c(0, 0, 0, 1, 0),
            c(0, 0, 0, 0, 1))
d <- Kt %*% coef(full)              # here m = 0
V <- Kt %*% vcov(full) %*% t(Kt)
F_star <- drop(t(d) %*% solve(V) %*% d) / 2
c(F_star, 1 - pf(F_star, 2, 385))
```

```
## [1] 7.6362147397 0.0005594001
```

$F^* = 7.64 > F(0.95;\, 2,\, 385) = 3.02$: reject H_0. Once weight and horsepower are in the model, displacement and cylinders together still explain some variation in acceleration.

Summary: testing coefficients

H_0	Statistic	Under H_0
$\boldsymbol{c}'\boldsymbol{\beta} = \theta_0$ (one combination)	$t^* = (\boldsymbol{c}'\hat{\boldsymbol{\beta}} - \theta_0)/s\{\boldsymbol{c}'\hat{\boldsymbol{\beta}}\}$	$t(n-p)$
$\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ (q restrictions)	F^* from $\boldsymbol{K}'\hat{\boldsymbol{\beta}} - \boldsymbol{m}$	$F(q,\, n-p)$

Extra sums of squares

The same test, from two fits

Compare the full model (F) with the reduced model (R) obtained by imposing H_0:

full (F) : `acc` $\sim$ `wt + hp + disp + cyl`, reduced (R) : `acc` $\sim$ `wt + hp`.

The same test, from two fits

Compare the full model (F) with the reduced model (R) obtained by imposing H_0:

full (F) : `acc` $\sim$ `wt + hp + disp + cyl`, reduced (R) : `acc` $\sim$ `wt + hp`.

The reduced model is a constrained version of the full one, so it fits less well: $SSE(R) \geq SSE(F)$. Is the increase large?

$$F^* = \frac{\big(SSE(R) - SSE(F)\big)/(df_R - df_F)}{SSE(F)/df_F}, \qquad df = \text{error degrees of freedom.}$$

The same test, from two fits

Compare the full model (F) with the reduced model (R) obtained by imposing H_0:

full (F) : `acc` $\sim$ `wt + hp + disp + cyl`, reduced (R) : `acc` $\sim$ `wt + hp`.

The reduced model is a constrained version of the full one, so it fits less well: $SSE(R) \geq SSE(F)$. Is the increase large?

$$F^* = \frac{\big(SSE(R) - SSE(F)\big)/(df_R - df_F)}{SSE(F)/df_F}, \qquad df = \text{error degrees of freedom.}$$

Here $df_R - df_F = q = 2$, and

$$F^* = \frac{(1174.3 - 1129.5)/2}{1129.5/385} = 7.64 : \quad \text{the same value as with } \boldsymbol{K}' \text{ in (F) only.}$$

The same test, from two fits

Compare the full model (F) with the reduced model (R) obtained by imposing H_0:

full (F) : `acc` $\sim$ `wt + hp + disp + cyl`, reduced (R) : `acc` $\sim$ `wt + hp`.

The reduced model is a constrained version of the full one, so it fits less well: $SSE(R) \geq SSE(F)$. Is the increase large?

$$F^* = \frac{\big(SSE(R) - SSE(F)\big)/(df_R - df_F)}{SSE(F)/df_F}, \qquad df = \text{error degrees of freedom.}$$

Here $df_R - df_F = q = 2$, and

$$F^* = \frac{(1174.3 - 1129.5)/2}{1129.5/385} = 7.64 : \quad \text{the same value as with } \boldsymbol{K}' \text{ in (F) only.}$$

The two forms of F^* always agree: fitting the model under H_0 is another way to compute it.

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, 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, 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, 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 precise.

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 precise.

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 precise.

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.

Summary

- The F distribution of the tests comes from the distribution of quadratic forms in normal vectors.
- One linear combination: a t test. Several restrictions $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$: an F test, which can also be computed from the SSE of a full and a reduced model.
- Extra sums of squares measure what predictors add once others are in; 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.

After reading week: qualitative predictors.