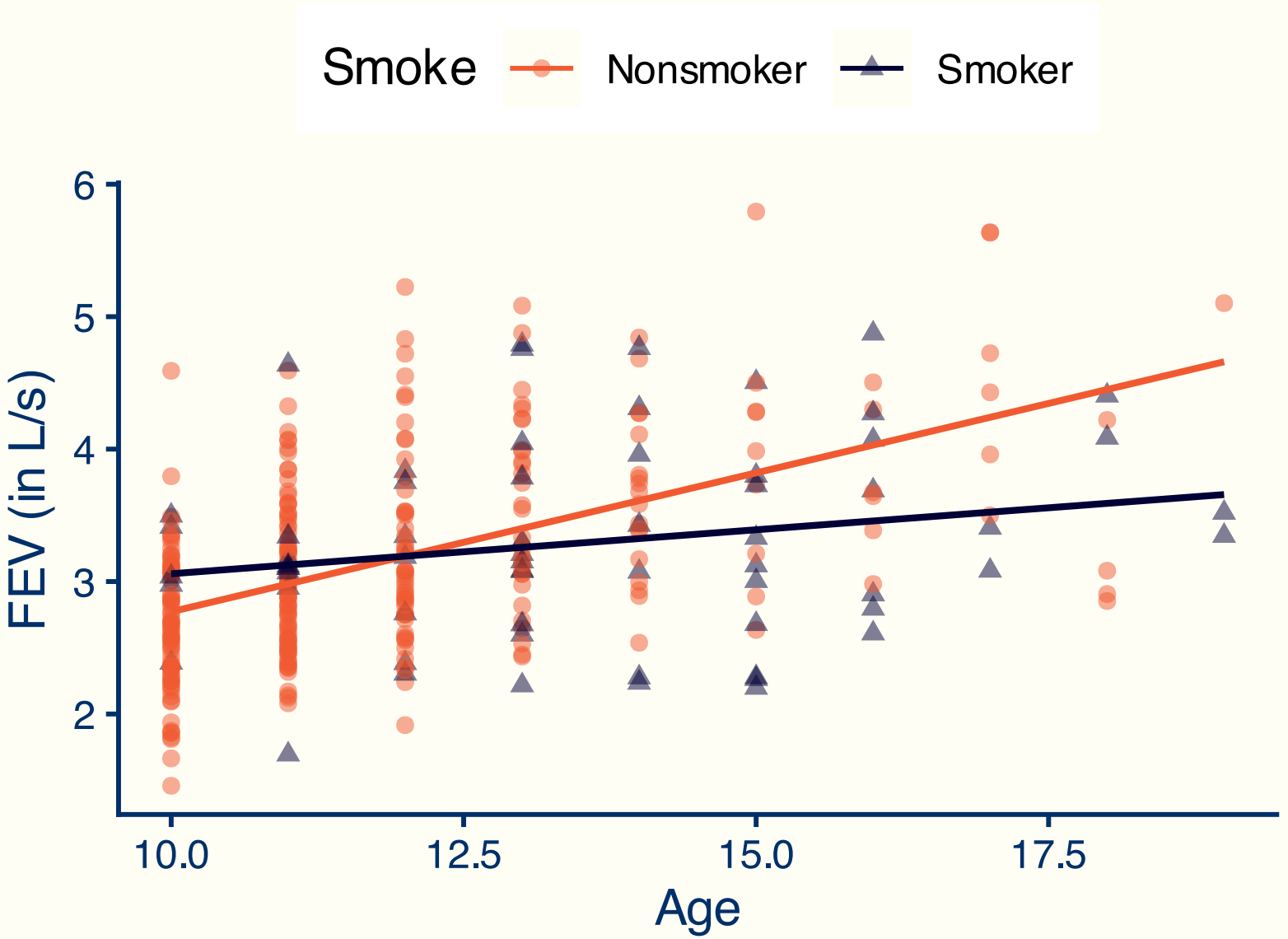
Additional Inferential Tools for MLR

Stat 230: Applied Regression Analysis

Linear Combinations

FEV example



FEV example

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{smoke} + \beta_3 \texttt{age} \times \texttt{smoke}$$

Mean functions for each group

Smokers: $\mathrm{E}(Y|X) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\texttt{age}$

Non-smokers: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 \texttt{age}$

SEs for linear combinations

Estimate:

$$\hat{\gamma} = c_1\hat{\beta}_i + c_2\hat{\beta}_j$$

Standard error:

$$SE(\hat{\gamma}) = \sqrt{c_1^2\{SE(\hat{\beta}_i)\}^2 + c_2^2\{SE(\hat{\beta}_j)\}^2 + 2c_1c_2Cov(\hat{\beta}_i, \hat{\beta}_j)}$$

FEV example

> ⚠ **Warning**
>
> Our standard table of coefficients doesn't contain the SEs we need to construct CIs or perform tests for these linear combinations.

term	estimate	std.error	statistic	p.value
(Intercept)	0.675	0.251	2.684	0.008
Age	0.210	0.021	9.972	<0.001
SmokeSmoker	1.720	0.563	3.053	0.002
Age:SmokeSmoker	−0.143	0.042	−3.397	<0.001

Covariance matrix

- The diagonal (top left to lower right) displays variances, $\{SE(\hat{\beta}_j)\}^2$
- The off-diagonals display covariances, $Cov(\hat{\beta}_i, \hat{\beta}_j)$
- Covariance is like correlation, but not scaled to be between -1 and 1

```
                (Intercept)         Age SmokeSmoker Age:SmokeSmoker
(Intercept)        0.063199 -0.0052210   -0.063199        0.0052210
Age               -0.005221  0.0004423    0.005221       -0.0004423
SmokeSmoker       -0.063199  0.0052210    0.317248       -0.0234032
Age:SmokeSmoker    0.005221 -0.0004423   -0.023403        0.0017798
```

SE for a linear combination

Smokers: $\mathrm{E}(Y|X) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\text{age}$

Target: $\hat{\gamma} = \hat{\beta}_1 + \hat{\beta}_3$

	(Intercept)	Age	SmokeSmoker	Age:SmokeSmoker
(Intercept)	0.063199	−0.0052210	−0.063199	0.0052210
Age	−0.005221	0.0004423	0.005221	−0.0004423
SmokeSmoker	−0.063199	0.0052210	0.317248	−0.0234032
Age:SmokeSmoker	0.005221	−0.0004423	−0.023403	0.0017798

$$SE(\hat{\gamma}) = \sqrt{c_1^2\{SE(\hat{\beta}_1)\}^2 + c_2^2\{SE(\hat{\beta}_2)\}^2 + 2c_1c_2Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

$c_1 = 1, c_2 = 1$

$SE(\widehat{\beta_1})^2 = 0.0004423$ $\qquad$ $SE(\widehat{\beta_2})^2 = 0.0017798$

$Cov(\widehat{\beta_1}, \widehat{\beta_2}) = -0.0004423$

CI for a linear combination

Once we have $SE(\hat{\gamma})$, we can construct a CI as usual:

$$\hat{\gamma} \pm t^*_{n-(p+1)} \cdot SE(\hat{\gamma})$$

$\hat{\gamma} = 0.0664$

$SE(\hat{\gamma}) = 0.0366$

$df = n - (p + 1) = 345 - 4 = 341$

$t^*_{341} \approx 1.649$ for a 90% CI

Test for a linear combination

We can also test hypotheses about linear combinations of coefficients using a t-test:

$$H_0 : \gamma = \gamma_0 \text{ vs. } H_a : \gamma \begin{matrix} < \\ \neq \\ > \end{matrix} \gamma_0$$

$$t = \frac{\hat{\gamma} - \gamma_0}{SE(\hat{\gamma})}$$

Reference distribution: $t_{n-(p+1)}$

Example

- Work through the example on the handout with your neighbors
- Be prepared to share your strategy with the class

Extra Sums of Squares F-tests

Example: Used car prices

We have data on used car prices, including:

- `Price`: Price (in dollars)
- `Mileage`: Mileage (in thousands of miles)
- `Make`: Manufacturer (Buick, Cadillac, Chevrolet, Pontiac, SAAB, Saturn)

After some initial EDA, researchers want to fit an MLR model to predict log(price) using mileage and make.

Used car prices

Is `Make` a useful predictor of price, after accounting for `Mileage`?

term	estimate	std.error	statistic	p.value
(Intercept)	10.090	0.034	293.975	<0.001
Mileage	0.000	0.000	−7.359	<0.001
MakeCadillac	0.649	0.038	17.016	<0.001
MakeChevrolet	−0.296	0.030	−9.809	<0.001
MakePontiac	−0.144	0.033	−4.299	<0.001
MakeSAAB	0.355	0.035	10.108	<0.001
MakeSaturn	−0.397	0.041	−9.650	<0.001

Review: Nested models

Model 1:

$$\mathrm{E}(\log(Y)|X) = \beta_0 + \beta_1 \texttt{Mileage}$$

Model 2:

$$\begin{aligned}\mathrm{E}(\log(Y)|X) = \beta_0 &+ \beta_1 \texttt{Mileage} + \beta_2 \texttt{Cadillac} + \beta_3 \texttt{Chevy} \\ &+ \beta_4 \texttt{Pontiac} + \beta_5 \texttt{SAAB} + \beta_6 \texttt{Saturn}\end{aligned}$$

Extra SS F-test:

Null hypothesis tells us how we can simplify Model 2 (full) to get Model 1 (reduced):

$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

Extra Sums of Squares

Definition: the marginal reduction in the error sum of squares when one (or more) explanatory variable(s) is added to the model

F-test for comparing nested models:

$$F = \frac{(SSE_{full} - SSE_{reduced})/(df_{full} - df_{reduced})}{MSE_{full}}$$

ANOVA table from R

`anova(car_lm)` produces an ANOVA table for the MLR model, it gives **sequential sums of squares**, not the extra sums of squares we need for our F-test.

term	df	sumsq	meansq	statistic	p.value
Mileage	1	2.959	2.959	50.941	<0.001
Make	5	85.802	17.160	295.414	<0.001
Residuals	797	46.297	0.058	NA	NA

- `Mileage` row: extra SS for `Mileage` (compared to intercept-only model)
- `Make` row: extra SS for `Make` after accounting for `Mileage`
- `Residuals` row: SSE for the full model