Wooldridge Example

- Model the salary of major league baseball players.

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

- Does performance have an effect on salary?
- Formulate as joint null hypothesis: $H_0 : \beta_3 = 0 , \ \beta_4 = 0 , \ \beta_5 = 0$
    - Economists might call this **exclusion restriction**—testing whether variables could be excluded from model.
- Are the three performance indicators associated with a change in salary?

Regression Output

$$\hat{\log}(salary) = \underset{(0.29)}{11.19} + \underset{(.0121)}{.0689} years + \underset{(.0026)}{.0126} gamesyr + \underset{(.00110)}{.00098} bavg + \underset{(.0161)}{.0144} hrunsyr + \underset{(.0072)}{.0108} rbisyr$$

$n = 353, \ \ \mathrm{SSR} = 183.186, \ \ \mathrm{R}^2 = .6278$

- How can we test these coefficients jointly?
    - Remove from regression, see how much worse model fit is.
    - Model fit: sum of squared residuals (SSR)

How Does the Model Fit?

Restricted model: performance variables removed

$$\hat{\log}\left(salary\right) = \underset{(0.11)}{11.22} + \underset{(.0125)}{.0713}years + \underset{(.0013)}{.0202}gamesyr$$

$n = 353, \ \ \mathrm{SSR} = 198.311, \ \ \mathrm{R}_2 = .5971$

- Taking out variables can only make the fit worse.
- Is the increase statistically significant?

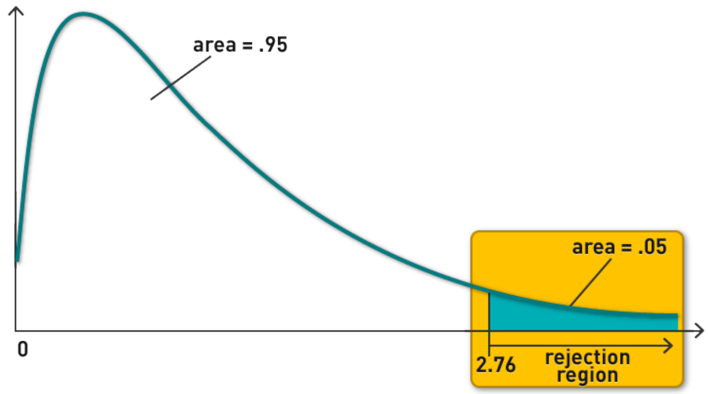
Forming a Test Statistic

$$F = \frac{\left(SSR_r - SSR_{ur}\right)/q}{SSR_{ur}/\left(n-k-1\right)} \sim F_{q,\,\mathrm{n}-k-1}$$

- Denominator: unrestricted SSR, divided by degrees of freedom in the unrestricted model
- Numerator: the change in SSR, divided by the number of variables being dropped
- Measuring the relative change in SSR, with constant scaling factors
- Under the null hypothesis, and assuming the CLM assumptions (MLR.1–6), test statistic follows an ***F*-distribution**.

F-Distribution

- To identify single distribution, specify both degrees of freedom for numerator and denominator.



- *F*-distribution only takes on positive values, corresponding to SSR only increasing when variables are removed. (*F*-statistic numerator will be positive.)
- The bigger the increase in SSR, the bigger our *F*-statistic and the further to the right of the distribution.
- Choose the critical value so that the null hypothesis is rejected in 5% of the cases, assuming it is true; in this case: 2.76.

F-Statistic

$$F = \frac{(198.311 - 183.186)/3}{183.186/(353-5-1)} \approx 9.55$$

$$F \sim F_{3,347} \Rightarrow c_{0.01} = 3.78$$

$$\Pr(F > 9.55) = 4.48 \times 10^{-6}$$

- Null hypothesis can be rejected, even at the 0.001 level.
- Variables are **jointly significant**.
- Variables were not significant when tested individually.
    - Likely reason: multicollinearity between them
    - Performance metrics tend to move up and down together; not much unique variation for OLS to work with.

Model Significance

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u$$

Application of joint significance: testing regression model as a whole

- **Omnibus test:** Can we exclude every *x* variable at the same time?
- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
- Restricted model is the mean: $y = \beta_0 + u$
- SSR: total sum of squares
- Does the model have any predictive power on the whole?

Model Significance (cont.)

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

- Test of overall significance is reported automatically in R.
- Most of the time, null hypothesis is automatically rejected.
    - If null can't be rejected, we may have little data or chosen variables lacking predictive power.
    - Model may be nonsignificant but can have a coefficient with a significant *t*-statistic.