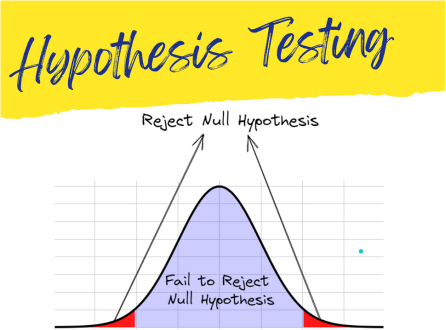
Fundamentals of Econometrics

Lecture 4: Multiple Linear Regression Model: Inference



Section 1

Multiple Regression Analysis: Inference

Review

Assumption	Result
MLR1. Specify true model MLR2. Data are random sample MLR3. No perfect collinearity MLR4. Zero conditional mean	OLS estimator is unbiased
MLR5. Homoskedasticity	OLS estimator is BLUE

Potential Problems discussed so far:

1. Omitted Variable Bias (MLR4. fails)
2. Multicollinearity

Hedonic Housing Price Model

- Goods are often treated as "homogenous" in economics.
  - What does this mean?
  - Is this a good assumption?

Hedonic models:

- Assume that people derive utility from the characteristics of goods or products.
- In equilibrium, therefore, the price of a good should reflect the value of its characteristics.
- Can use OLS to estimate the value (implicit prices) of these characteristics.

Example: Hedonic Housing Price Model

Suppose we want to estimate the environmental impact of agricultural externalities on housing prices in San Joaquin, CA.

- Grazing land provides a scenic view and open spaces, but may also attract pests.
- Crop production may generate noise and dust, and health concerns from pesticide use.

Data

- salesprice = sales price of house in San Joaquin, CA in 1998
- gdistance = distance in meters to nearest grazing land
- wdistance = distance in meters to nearest wetland
- cdistance = distance in meters to nearest cropland
- bathrooms = number of bathrooms
- bedrooms = number of bedrooms
- sqftbuilding = square feet of building
- sqftlot = square feet of lot
- age = age of home

```r
library(foreign) # for reading Stata files
sanjoaquin <- read.dta("../../data/San_Joaquin.dta")
stargazer(sanjoaquin, font.size="footnotesize",
          header = FALSE, title = "Descriptive Statistics")
```

Table 1: Descriptive Statistics

Statistic	N	Mean	St. Dev.	Min	Max
saleprice	2,661	130,072.000	52,067.000	30,000	355,000
gdistance	2,661	8,342.000	4,401.000	11.100	15,940.000
wdistance	2,661	7,312.000	5,148.000	3.030	26,788.000
cdistance	2,661	886.000	778.000	0.152	3,472.000
bathrooms	2,661	1.900	0.605	1.000	4.500
bedrooms	2,661	3.050	0.705	1	6
sqftbuilding	2,661	1,533.000	499.000	366	4,096
sqftlot	2,661	8,669.000	12,231.000	1,300	217,800
age	2,661	24.900	21.200	1	98

```r
summary(hedonic <- lm(saleprice ~ ., data = sanjoaquin))
```

```
##
## Call:
## lm(formula = saleprice ~ ., data = sanjoaquin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -255118  -16289   -1536   14753  239339
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.63e+04   4.31e+03   10.75  < 2e-16 ***
## gdistance    -1.61e+00   1.77e-01   -9.10  < 2e-16 ***
## wdistance     6.65e-01   1.59e-01    4.19  2.9e-05 ***
## cdistance     2.44e+00   9.29e-01    2.63   0.0087 **
## bathrooms     2.46e+03   1.76e+03    1.40   0.1621
## bedrooms     -5.86e+03   1.16e+03   -5.03  5.2e-07 ***
## sqftbuilding  7.28e+01   1.94e+00   37.56  < 2e-16 ***
## sqftlot       5.06e-01   4.85e-02   10.44  < 2e-16 ***
## age          -5.06e+02   3.73e+01  -13.54  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29600 on 2652 degrees of freedom
## Multiple R-squared:  0.678,  Adjusted R-squared:  0.677
## F-statistic:  696 on 8 and 2652 DF,  p-value: <2e-16
```

Interpretations

Holding all other independent variables constant:

- **Distance to grazing land:** The sales price for a home decreases by \$1.609 for every meter we move away from the nearest grazing land.
- **Distance to nearest cropland:** The sales price for a home increases by \$2.44 for every meter we move away from the nearest cropland.
- **Bedrooms:** The sales price for a home decreases by \$5855.396 for every additional bedroom.
  - Does this make sense?
    - Since all other independent variables are held constant (including square footage of the house), more bedroom would imply a smaller size of each bedroom (thus lower price).
    - A better specification would be to interact bedrooms with `sqftbuilding`.

Distribution of OLS Estimators

- Our OLS estimators depend on the error term,u, and by extension, the distribution of u.
- For statistical testing, we need to know the sampling distributions of the OLS estimators.
- MLR6. Population error (u) is independent of the explanatory variables, $x_1, x_2, \ldots, x_k$, and normally distributed with zero mean and variance σ^2.: $u \sim N(0, \sigma^2)$
- MLR 1-6 are called the **Classical Linear Model assumptions**.

Is Normality a strong assumption?

- MLR6. Implies that MLR4 and MLR5 hold.
  - In sample size is small, MLR6 can be very strong and just as important as the conditional mean assumption.
  - It becomes increasingly less important as the sample size grows increasingly large.
  - If MLR6 holds, then our estimators will also be normally distributed

Normal Distributions

Recall that the normal distribution is

- symmetric around the mean
- has a bell-shaped curve
- Tail stretches to infinity

Some other properties of the normal distribution:

1. Any linear combination of independent identically distributed normal random variables is also normally distributed.
2. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$.

Normal Distribution

1. Any linear combination of independent identically distributed (*iid*) normal random variables is also normally distributed.

$$
\begin{aligned}
x_i \overset{iid}{\sim} N(\mu, \sigma^2) &\quad \omega = x_1 + 2x_2 - 3x_3 \\
E(\omega) =& E(x_1) + 2E(x_2) - 3E(x_3) = \mu + 2\mu - 3\mu = 0 \\
\text{var}(\omega) =& \text{var}(x_1) + 4\text{var}(x_2) + 9\text{var}(x_3) = \sigma^2 + 4\sigma^2 + 9\sigma^2 = 14\sigma^2 \\
&\implies \omega \sim N(0, 14\sigma^2)
\end{aligned}
$$

2. If $X \overset{iid}{\sim} N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

$$
\begin{aligned}
E\left[\frac{x-\mu}{\sigma}\right] &= \frac{\cancelto{\mu}{E(x)} - \mu}{\sigma} = 0 \\
var\left(\frac{x-\mu}{\sigma}\right) &= \frac{\text{var}(x-\mu)}{\sigma^2} = \frac{\sigma^2 - 0}{\sigma^2} = 1
\end{aligned}
$$

Distribution of OLS Estimators

Recall from our earlier discussions that:
$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) = \boxed{\beta + (X'X)^{-1}X'u}$

By MLR6 and the first property of the Normal distribution:

$$\hat{\beta}_j \sim N\left[\beta_j, var(\hat{\beta}_j)\right]$$

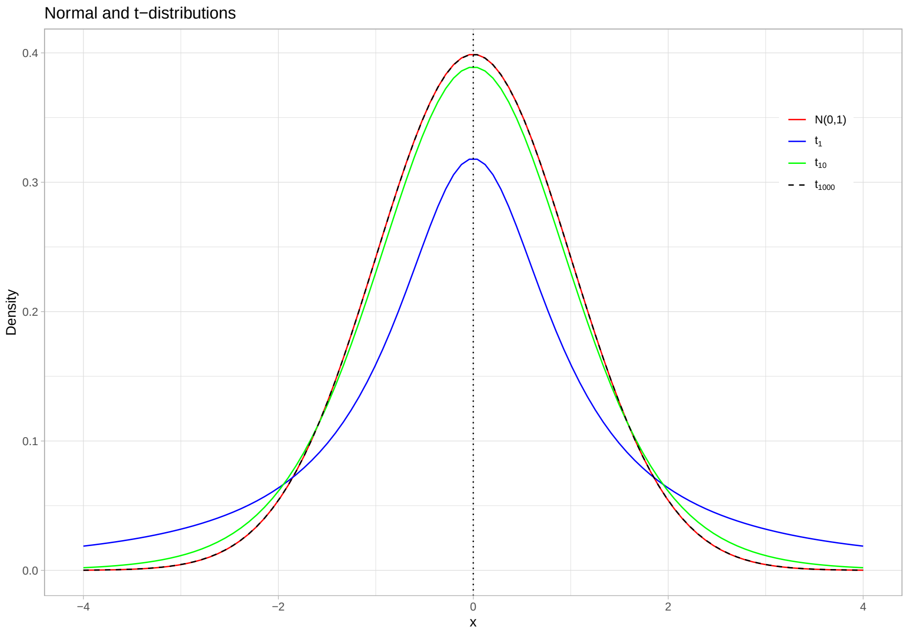
By the second property of the Normal distribution:

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$$

For hypothesis testing therefore, we use

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

where t_{n-k-1} is the students t-distribution with $n-k-1$ degrees of freedom.


Normal and t−distributions
Density
x
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
N(0,1)
t_1
t_{10}
t_{1000}

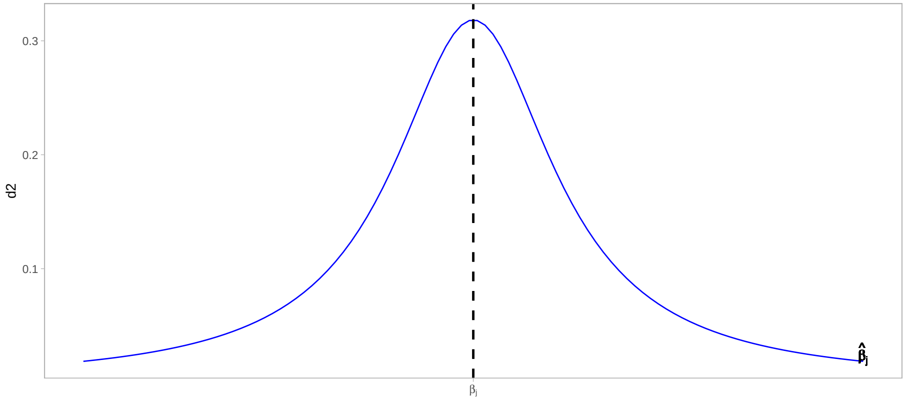
Section 2

Single Parameter Hypothesis Testing

Hypothesis Testing

- Why do we do hypothesis testing?
  - We might want to be able to make statements about the probability of observing a certain outcome (or value of $\hat{\beta}$).
- If MLR1-MLR4 hold, we know that our estimate of β is unbiased.
- **But for any given random sample, the actual estimate may be anywhere along the distribution of $\hat{\beta}$.** Think back to our Monte Carlo simulation exercises.
- The question is: **How do we know whether the estimate we have is "close enough" to some hypothesized value of β?**

Hypothesis Testing



Potential Question

- How likely is it that the true value of β_j is equal to 0?

One-Sided Hypothesis Testing

1. Hypothesis

Null Hypothesis:	$H_0 : \beta_j = 0$ (or $\beta_j \leq 0$)
Alternative Hypothesis:	$H_1 : \beta_j > 0$ (or $\beta_j < 0$)

2. Test Statistic

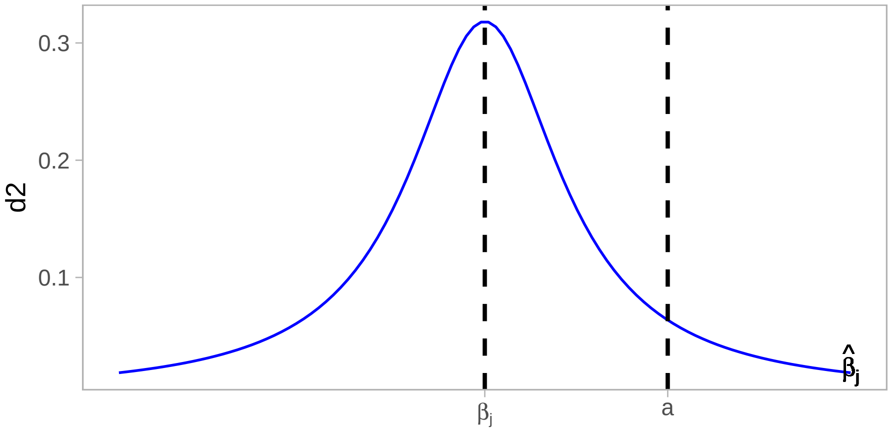
Our test statistics under the null hypothesis is:

$$t_{stat} = \frac{\hat{\beta}_j - 0}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

3. Decision Rule

We **reject** the null hypothesis if $t_{stat} > t_{n-k-1,\alpha}$, otherwise, we **fail to reject**.

Here $t_{n-k-1,\alpha}$ is the critical value of the t-distribution with $n - k - 1$ degrees of freedom at significance level α.



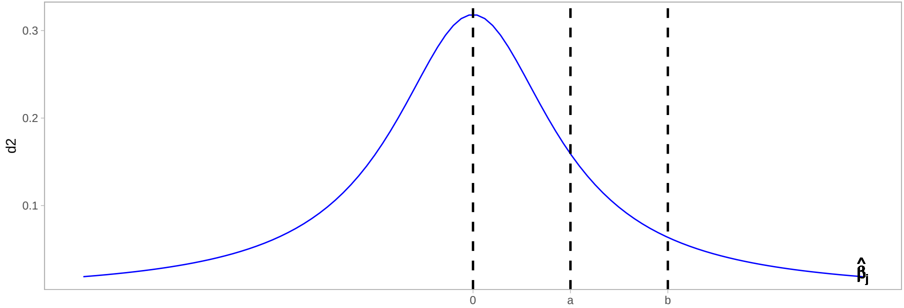
Why is the distribution of $\hat{\beta}_j$ important?

- We want to be able to make statements about the probability of observing a certain outcome (or value of $\hat{\beta}$).

For example, how likely would it be to observe a value of $\hat{\beta}_j \geq a$?

(Another) Graphical Illustration

Assume the following distribution for the t-statistic under the null:



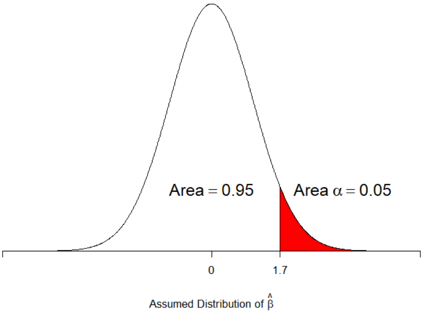
- At point `b`, we are more likely to reject than at point `a`.
- The basic question is "how do we know whether point `a` or point `b` is large enough to reject the null hypothesis?"

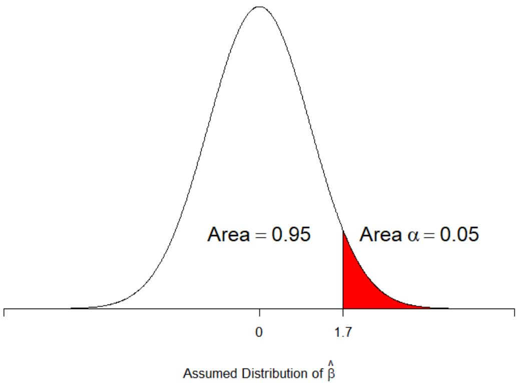
Critical Value

In Hypothesis testing, we can make 2 types of mistakes:

1. **Type I Error**: Rejecting the null hypothesis when it is true.
2. **Type II Error**: Failing to reject the null hypothesis when it is false.

- Our critical values are chosen to make the probability of making a Type I error small.
- We can control this probability by setting a significance level, α.





- For a t-distribution with $n - k - 1 = 28$ degrees of freedom, a t-value of 1.701 corresponds to a 5% probability of making a Type I error (1-tail).
- The probability of making a Type I error is 0.01 if the t-value is 2.462 (one-tailed).

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Finding Critical values in R

```r
# Critical value (t-crit) of t(28,0.05), one-tailed
qt(p = 0.05, df = 28)
```

```
## [1] -1.7
```

```r
# Critical value (t-crit) of t(28,0.01), one-tailed
qt(p = 0.01, df = 28)
```

```
## [1] -2.47
```

Does lot sizes increase the price of a house?

1. Hypothesis

$$H_0 : \beta_{sqftlot} = 0$$
$$H_1 : \beta_{sqftlot} > 0$$

2. Test Statistic

$$t_{stat} = \frac{\hat{\beta}_{sqftlot}}{se(\hat{\beta}_{sqftlot})} \sim t_{n-k-1} = \frac{0.506}{0.048} = 10.444$$

3. Decision Rule: Reject H_0 if $t_{stat} > t_{n-k-1,\alpha}$, otherwise, fail to reject.

What else do we need?

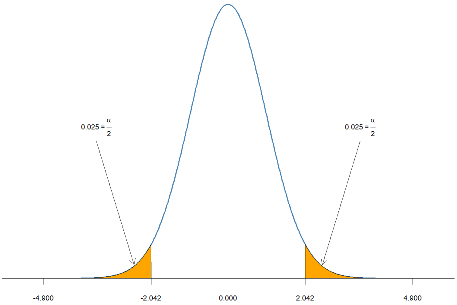
- Level of significance: α.
- dof, $n - k - 1$:(2652)
- $t_{n-k-1,\alpha}$.

Two-sided Hypothesis

Economic theory may not tell us what the sign of the coefficient should be. Instead, we may be interested in whether x has any effect on y.

1. Hypothesis

$$H_0 : \beta_j = 0$$
$$H_1 : \beta_j \neq 0$$



Reject H_0 **if** $|t_{stat}| > t_{n-k-1,\alpha/2}$

Does the number of bathrooms affect the price?

1. Hypothesis

$$H_0 : \beta_{bathrooms} = 0$$
$$H_1 : \beta_{bathrooms} \neq 0$$

2. Test Statistic

$$t_{stat} = \frac{2459.875}{1759.153} = 1.398$$

3. Decision Rule: Reject H_0 if $|t_{stat}| > t_{n-k-1,\alpha/2}$, otherwise, fail to reject.

Critical value: $t_{2652,0.025} = 1.961$.

4. Conclusion: Since $|1.398| < 1.961$, we **fail to reject** the null hypothesis and conclude that at the 5% level of significance, the number of bathrooms in a house does not affect its price.

What about the number of bedrooms?

1. Hypothesis

2. Test Statistic

3. Decision Rule

Critical value:

4. Conclusion:

Testing for a Specific Value

- Sometimes, we may want to test for a specific value of β_j.
  - Here, a value other than zero may be of interest.
  - These tests could be one-sided or two-sided.
- The test statistic is the same as before:

$$t_{stat} = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

Here β_j is the hypothesized value against which we are testing.

Annual Crimes on college campuses

Suppose we are interested in testing whether the growth rate of annual crimes on college campuses **is proportional** to the growth rate of student enrollment.

Using the `campus` dataset, we estimate the following model:

$$log(crimes) = \beta_0 + \beta_1 log(enroll) + u$$

```
(lm(lcrime ~ lenroll, data = campus) |> summary())[c(4,7)]
```

```
## $coefficients
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -6.63       1.03   -6.42 5.44e-09
## lenroll         1.27       0.11   11.57 7.83e-20
##
## $df
## [1]  2 95  2
```

Annual Crimes on college campuses

1. Hypothesis

2. Test Statistic

3. Decision Rule

Critical value:

4. Conclusion:

Annual Crimes on college campuses

How would the test look different if we were interested in testing whether the growth rate of annual crimes on college campuses **is more than proportional** to the growth rate of student enrollment?

Are there potential problems with this model?

- We have not controlled for other factors that may affect the number of crimes on college campuses.
- Is the college campus located in a high-crime area? Urban or rural?

Confidence Intervals

- Hypothesis testing is a binary decision: reject or fail to reject.
- Confidence intervals provide a range of values within which we are confident the true value of β_j lies.
- The confidence interval is constructed as:

$$P\left(\underbrace{\hat{\beta}_j - c_{\alpha/2} \cdot se(\hat{\beta}_j)}_{\text{Lower bound of CI}} \leq \beta_j \leq (\underbrace{\hat{\beta}_j + c_{\alpha/2} \cdot se(\hat{\beta}_j)}_{\text{Upper bound of CI}}\right) = 1 - \alpha$$

where $c_{\alpha/2}$ is the **critical value** of the two-sided test and $1 - \alpha$ is the **confidence level**.

Interpretation of the Confidence Interval

- The bounds of the confidence interval are random.
- If we were to repeat the experiment many times, we would expect the true value of β_j to lie within the confidence interval in $(1 - \alpha)\%$ of the experiments.

Confidence Intervals

Typical Confidence Levels

$$P\left(\hat{\beta}_j - c_{0.01/2} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.01/2} \cdot se(\hat{\beta}_j)\right) = 0.99$$
$$P\left(\hat{\beta}_j - c_{0.05/2} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.05} \cdot se(\hat{\beta}_j)\right) = 0.95$$
$$P\left(\hat{\beta}_j - c_{0.10/2} \cdot se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{0.10} \cdot se(\hat{\beta}_j)\right) = 0.90$$

- Use rule of thumb: $c_{0.01/2} = 2.58$, $c_{0.05/2} = 1.96$, $c_{0.10/2} = 1.645$.

Relationship between Confidence Intervals and Hypothesis Testing

$$a_j \notin CI \implies H_0 : \beta_j = a_j \text{ is rejected}$$

- If the confidence interval does not contain the hypothesized value, then we would reject the null hypothesis at the α level of significance.

```r
gpa.mod <- lm(colGPA~ hsGPA + ACT + skipped, data = gpa1)
gpa.mod |> summary()
```

```
##
## Call:
## lm(formula = colGPA ~ hsGPA + ACT + skipped, data = gpa1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8570 -0.2320 -0.0393  0.2482  0.8166
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3896     0.3316    4.19  5.0e-05 ***
## hsGPA         0.4118     0.0937    4.40  2.2e-05 ***
## ACT           0.0147     0.0106    1.39   0.1658
## skipped      -0.0831     0.0260   -3.20   0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.329 on 137 degrees of freedom
## Multiple R-squared:  0.234,  Adjusted R-squared:  0.217
## F-statistic: 13.9 on 3 and 137 DF,  p-value: 5.65e-08
```

Confidence Intervals

$$\widehat{colGPA} = \underset{(0.332)}{1.390} + \underset{(0.094)}{0.412} hsGPA + \underset{(0.011)}{0.015} ACT + \underset{(0.026)}{-0.083} skipped$$

Are ACT scores significantly related to college GPA?

$$H_0 : \beta_{ACT} = 0$$
$$H_1 : \beta_{ACT} \neq 0$$

df : $n - k - 1 = 137$, $c_{0.1/2} = 1.656$

$$\hat{\beta}_{ACT} \pm c_{0.1/2} \cdot se(\hat{\beta}_{ACT}) \implies 0.015 \pm 0.017 = (-0.003, 0.032)$$

Confidence Intervals

```r
# CI for all parms
gpa.mod |> confint(level = 0.90)
```

```
##                  5 %    95 %
## (Intercept)  0.84048  1.9386
## hsGPA        0.25669  0.5669
## ACT         -0.00278  0.0322
## skipped     -0.12617 -0.0401
```

```r
# CI for ACT only at 95% level
gpa.mod |> confint(parm = "ACT", level = 0.95)
```

```
##         2.5 % 97.5 %
## ACT -0.00617 0.0356
```

Computing `p-Values` for t-tests.

- Rather than testing at every significance level, we might find it more convenient to ask: "Given the observed `t-stat` value, what is the *smallest* significance level at which we would reject the null hypothesis?"
- The `p-value` tells the strength (or weakness) of the evidence against the null hypothesis.
- In short, the `p-value` is the **probability of observing a value of the test statistic as extreme as the one observed, given that the null hypothesis is true**.
- The smaller the `p-value`, the stronger the evidence against the null hypothesis.

Computing p-Values for t-tests.

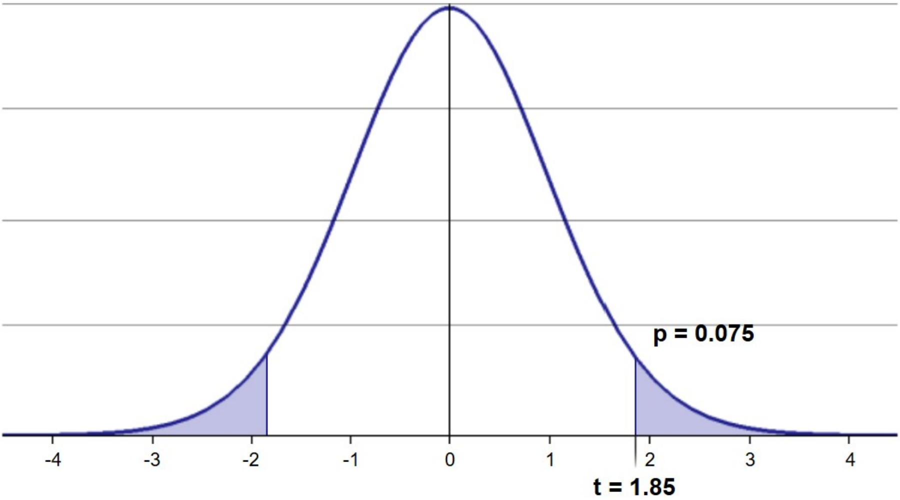
Example Explained

```
2*(1-pt(q = 1.85, df = 28))
```

```
## [1] 0.0749
```

- This means that, if the null hypothesis were true, we would expect to observe an **absolute value** of the test statistic as extreme as 1.85 about 7.489% of the time.
- This provides some evidence against the null hypothesis but we would not reject the null hypothesis at the 5% level of significance.

What about at the 10% level of significance?


p = 0.075
-4
-3
-2
-1
0
1
2
3
4
t = 1.85

```r
# prob of a t-crit = 1.701 and df = 28, one tailed
1-pt(q = 1.701, df = 28)
```

```
## [1] 0.05
```

```r
# prob of a t-crit = 2.462 and df = 28, one tailed
1-pt(q = 2.462, df = 28)
```

```
## [1] 0.0101
```

```r
# prob of a t-crit = 2.462 and df = 28, two tailed
2*(1-pt(q = 2.462, df = 28))
```

```
## [1] 0.0202
```

Section 3

Multiple Parameter Hypothesis Testing

Single Linear Combinations of Parameters

- Sometimes, we might want to a single hypothesis test on more than one parameter.
  - For example, do people attending a junior college have the same returns to education as those attending a 4-year college? `data(twoyear)`

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

    where `jc` = total 2-year college credits, `univ` = total 4-year college credits, `exper` = total (actual) work experience.
- Under the null, we are interest in whether on year at a junior college is worth one year at a 4-year college:

$$H_0 : \beta_1 = \beta_2$$

- We can assume a one-sided alternative that a year of junior college is worth less than a year at a 4-year college:

$$H_1 : \beta_1 < \beta_2$$

Single Linear Combinations of Parameters

- Given the problem above, we cannot simple conduct two separate hypothesis tests on β_1 and β_2.
- Instead, we can rewrithe the null and alternative hypotheses to text the following:

$$H_0 : \beta_1 - \beta_2 = 0$$
$$H_1 : \beta_1 - \beta_2 < 0$$

- The t-test is now based on whether the difference in the parameters is significantly different from zero.
- For ease of notation and generality to problems later on, we can let $\beta_1 - \beta_2 = \theta$, such that:

$$H_0 : \theta = 0$$
$$H_1 : \theta < 0$$

Single Linear Combinations of Parameters

The test statistic is given by:

$$t_{stat} = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{\hat{\theta} - 0}{se(\hat{\theta})} \sim t_{n-k-1}$$

- Recall that the standard error of $se(\hat{\beta}_1 - \hat{\beta}_2)$ is given by $\sqrt{var(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2\text{cov}(\hat{\beta}_1, \hat{\beta}_2)}$.

```r
(two.year <- lm(lwage ~ jc + univ + exper, data = twoyear)) |> summary()
```

```
##
## Call:
## lm(formula = lwage ~ jc + univ + exper, data = twoyear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1036 -0.2813  0.0055  0.2852  1.7817
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.472326   0.021060   69.91   <2e-16 ***
## jc          0.066697   0.006829    9.77   <2e-16 ***
## univ        0.076876   0.002309   33.30   <2e-16 ***
## exper       0.004944   0.000157   31.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.43 on 6759 degrees of freedom
## Multiple R-squared:  0.222,  Adjusted R-squared:  0.222
## F-statistic:  645 on 3 and 6759 DF,  p-value: <2e-16
```

```r
(vcov.two.year <- vcov(two.year))
```

```
##               (Intercept)        jc      univ     exper
## (Intercept)      4.44e-04 -1.74e-05 -1.57e-05 -3.10e-06
## jc              -1.74e-05  4.66e-05  1.93e-06 -1.72e-08
## univ            -1.57e-05  1.93e-06  5.33e-06  3.93e-08
## exper           -3.10e-06 -1.72e-08  3.93e-08  2.48e-08
```

t-stat then is:

$$t_{stat} = \frac{0.067 - 0.077}{\sqrt{0.00005 + 0.00001 - 2 * 0.00000}} = -1.468$$

Alternate Approach

- Recall, we said we could redefine $\theta = \beta_1 - \beta_2$. We can redefine the model as:

$$\begin{aligned} log(wage) &= \beta_0 + (\theta + \beta_2)jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta jc + \beta_2(jc + univ) + \beta_3 exper + u \end{aligned}$$

The null of $\beta_1 = \beta_2$ is now equivalent to testing whether $\theta = 0$, as we noted earlier.

$$H_0 : \theta = 0$$
$$H_1 : \theta < 0$$

```r
(trans.mod <- lm(lwage ~ jc + I(jc+univ) + exper, data = twoyear)) |>
  summary()
```

```
##
## Call:
## lm(formula = lwage ~ jc + I(jc + univ) + exper, data = twoyear)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1036 -0.2813  0.0055  0.2852  1.7817
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.472326   0.021060   69.91   <2e-16 ***
## jc           -0.010180   0.006936   -1.47     0.14
## I(jc + univ)  0.076876   0.002309   33.30   <2e-16 ***
## exper         0.004944   0.000157   31.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.43 on 6759 degrees of freedom
## Multiple R-squared:  0.222,  Adjusted R-squared:  0.222
## F-statistic:  645 on 3 and 6759 DF,  p-value: <2e-16
```

```r
# 90%, 95%, and 99% CI for theta
levelss <- c("90%" = .9, "95%" =.95, "99%" = .99)
lapply(levelss, function(x){confint(trans.mod, level = x)})
```

```
## $`90%`
##                     5 %    95 %
## (Intercept)     1.43768 1.50697
## jc             -0.02159 0.00123
## I(jc + univ)    0.07308 0.08067
## exper           0.00469 0.00520
##
## $`95%`
##                   2.5 %  97.5 %
## (Intercept)     1.43104 1.51361
## jc             -0.02378 0.00342
## I(jc + univ)    0.07235 0.08140
## exper           0.00464 0.00525
##
## $`99%`
##                   0.5 %  99.5 %
## (Intercept)     1.41806 1.52659
## jc             -0.02805 0.00769
## I(jc + univ)    0.07093 0.08282
## exper           0.00454 0.00535
```

Testing Multiple Restrictions

- We often want to test multiple restrictions on the parameters of a model.
- For example, we may want to test whether the coefficients on `jc` and `univ` are equal to each other and also equal to zero.
  - In other words, are they **jointly** insignificant?
  - That is, these set of exogenous variables have no partial effect (no explanatory power) on the dependent variable.
  - And if so, we can drop them from the model.

Example: Major League Baseball Salaries

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyrs + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u \quad (1)$$

where the variables are as defined in the `mlb1` dataset.

- Suppose we wanted to test the hypothesis that, once years in the league and games per year have been controlled for, all performance measures– $bavg$, $hrunsyr$, and $rbisyr$– have no effect on salary.
- Essentially the null states that productivity measured by baseball statistics has no effect on salary.
- The null and alternative hypotheses are:

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$
$$H_1 : \text{at least one of the } \beta_j \text{ differs from zero}; j \in \{3, 4, 5\}$$

Exclusion Restrictions

- The standard t-test from earlier is not appropriate here because the t-test is designed for only a single parameter as it puts no restrictions on the other parameters.
- Since we will need to test these exclusion restrictions **jointly**, we will need to use the **F-test**.
- We need to understand about the restricted vs. unrestricted model in order to perform an F-test.

Restricted Model

This is the model implied by the null hypothesis (the model with the restrictions imposed):

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyrs + u \tag{2}$$

- Of course, (1) is the unrestricted model.

Exclusion Restrictions

- Since the restricted model is a direct subset of the unrestricted model, we say that both models are **nested**.
- For nested models, we can use the following F-test to test the null hypothesis:

$$F = \frac{(SSR_r - SSR_{ur})/(df_r - df_u)}{SSR_{ur}/\mathrm{df}_{ur}} = \frac{SSR_r - SSR_{ur}}{q} \Big/ \frac{SSR_{ur}}{n-k-1}$$
$$= \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)} \sim F(q, n-k-1)$$

where: q = number of restrictions imposed (in this case, $q = 3$), SSR_r = sum of squared residuals from the restricted model, SSR_{ur} = sum of squared residuals from the unrestricted model, and df_r and df_u are the degrees of freedom for the restricted and unrestricted models respectively.

Unrestricted Model: Baseball Salary

```r
(unrest.mlb <- lm(lsalary ~ years + gamesyr + bavg + hrunsyr + rbisyr,
                  data = mlb1)) |> summary()
```

```
##
## Call:
## lm(formula = lsalary ~ years + gamesyr + bavg + hrunsyr + rbisyr,
##     data = mlb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0251 -0.4503 -0.0401  0.4701  2.6892
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.12e+01   2.89e-01   38.75  < 2e-16 ***
## years       6.89e-02   1.21e-02    5.68  2.8e-08 ***
## gamesyr     1.26e-02   2.65e-03    4.74  3.1e-06 ***
## bavg        9.79e-04   1.10e-03    0.89     0.38
## hrunsyr     1.44e-02   1.61e-02    0.90     0.37
## rbisyr      1.08e-02   7.17e-03    1.50     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Restricted Model: Baseball Salary

```r
(rest.mlb <- lm(lsalary ~ years + gamesyr, data = mlb1)) |> summary()
```

```
##
## Call:
## lm(formula = lsalary ~ years + gamesyr, data = mlb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6686 -0.4641 -0.0118  0.4922  2.6883
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.22380    0.10831   103.6  < 2e-16 ***
## years        0.07132    0.01251     5.7  2.5e-08 ***
## gamesyr      0.02017    0.00134    15.0  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.753 on 350 degrees of freedom
## Multiple R-squared:  0.597,  Adjusted R-squared:  0.595
## F-statistic:  259 on 2 and 350 DF,  p-value: <2e-16
```

```r
# F-statistic
SSR_r <- sum(resid(rest.mlb)^2)
SSR_ur <- sum(resid(unrest.mlb)^2)
df_r <- rest.mlb$df.residual
df_ur <- unrest.mlb$df.residual
q <- df_r - df_ur
(Fstat <- (SSR_r - SSR_ur)/(df_r - df_ur) / (SSR_ur/df_ur))
```

```
## [1] 9.55
```

$$
\begin{aligned}
F - stat &= \frac{(SSR_r - SSR_{ur})/(df_r - df_u)}{SSR_{ur}/\mathrm{df}_{ur}} \\
&= \frac{15.125}{3} \Big/ \frac{183.186}{347} \\
&= 9.550
\end{aligned}
$$

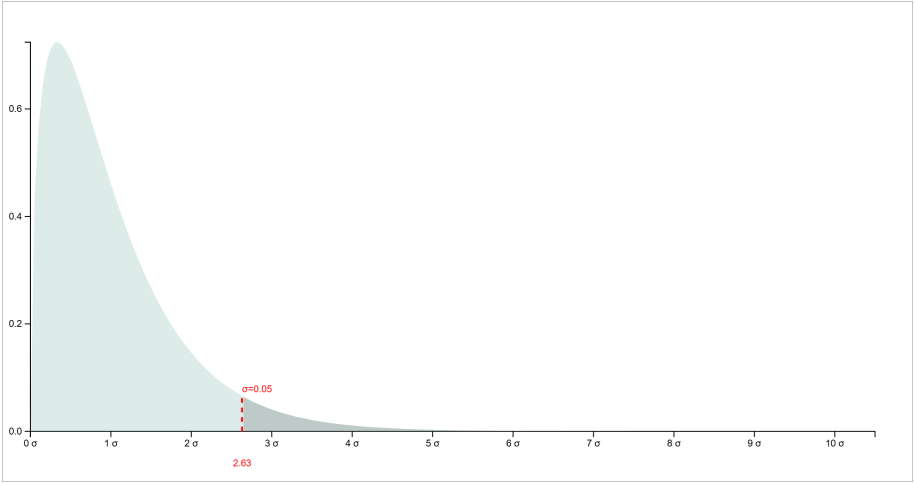
Critical Value

```r
# Critical value for F(3, 347) at 5% level of significance
qf(p = c("1%" = 0.01, "5%" = 0.05, "10%" = 0.1),
   df1 = 3, df2 = 347, lower.tail = FALSE)
```

```
##   1%   5%  10%
## 3.84 2.63 2.10
```

Conclusion:

Since our F-stat, $9.55 >$ the critical value of $F_{3,347,5\%} = 2.631$, we reject the null hypothesis and conclude that at the 5% level of significance, **at least 1** of the three statistics measuring performance– $bavg$, $hrunsyr$, and $rbisyr$– have a statistically significant effect on salary. That is, **they are jointly significant**.



0.6
0.4
0.2
0.0
σ=0.05
0 σ
1 σ
2 σ
3 σ
4 σ
5 σ
6 σ
7 σ
8 σ
9 σ
10 σ
2.63

Critical Values of the F-Distribution: $\alpha = 0.05$

Denom.	Numerator Degrees of Freedom									
d.f.	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145	2.098
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131	2.084
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
41	4.079	3.226	2.833	2.600	2.443	2.330	2.243	2.174	2.118	2.071
42	4.073	3.220	2.827	2.594	2.438	2.324	2.237	2.168	2.112	2.065
43	4.067	3.214	2.822	2.589	2.432	2.318	2.232	2.163	2.106	2.059
44	4.062	3.209	2.816	2.584	2.427	2.313	2.226	2.157	2.101	2.054
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096	2.049
46	4.052	3.200	2.807	2.574	2.417	2.304	2.216	2.147	2.091	2.044
47	4.047	3.195	2.802	2.570	2.413	2.299	2.212	2.143	2.086	2.039
48	4.043	3.191	2.798	2.565	2.409	2.295	2.207	2.138	2.082	2.035
49	4.038	3.187	2.794	2.561	2.404	2.290	2.203	2.134	2.077	2.030
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993

```r
car::linearHypothesis(unrest.mlb, c("bavg = 0", "hrunsyr = 0",
                                    "rbisyr = 0"))
```

```
##
## Linear hypothesis test:
## bavg = 0
## hrunsyr = 0
## rbisyr = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ years + gamesyr + bavg + hrunsyr + rbisyr
##
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1    350 198
## 2    347 183  3      15.1 9.55 4.5e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Individual vs Joint Significance

- Notice that the three variables that we restricted, $bavg$, $hrunsyr$, and $rbisyr$– are all insignificant at all conventional levels of significance.
- However, the F-test indicates that at least one of them is significant.
- This is a common problem when testing for joint significance, as the F-test masks the individual significance of the variables.
- The F-test focuses on the joint distribution of the parameters, while the t-test focuses on the marginal distribution of each parameter.

Overall Model Significance

What if we wanted to test whether the entire model is nonsensical and none of the parameters are significant?

- We can test the null hypothesis that all of the parameters are equal to zero:

$$H_0 : \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$
$$H_1 : \text{at least one of the } \beta_j \text{ differs from zero}; j \in \{1, 2, 3, 4, 5\}$$

```
car::linearHypothesis(unrest.mlb,
                      c("years = 0", "gamesyr = 0",
                        "bavg = 0",
                        "hrunsyr = 0", "rbisyr = 0"))
```

```
##
## Linear hypothesis test:
## years = 0
## gamesyr = 0
## bavg = 0
## hrunsyr = 0
## rbisyr = 0
##
## Model 1: restricted model
## Model 2: lsalary ~ years + gamesyr + bavg + hrunsyr + rbisy
##
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1    352 492
## 2    347 183  5       309 117 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

But where have we seen that before?

```r
# F-statistic of the overall model significance
(fmod <- summary(unrest.mlb)$fstat)
```

```
## value numdf dendf
##   117     5   347
```

```r
# P-value of the overall model significance
pf(fmod["value"], df1 = fmod["numdf"],
   df2 = fmod["dendf"], lower.tail = FALSE)
```

```
##    value
## 2.94e-72
```

If we returned to the R^2 approach, we would find that the F-statistic is equivalent to:

$$F = \frac{(R^2_{ur} - R^2_r)/[(n-k-1)-(n-1)]}{(1-R^2_{ur})/(n-k-1)} = \frac{R^2_{ur}/k}{(1-R^2_{ur})/(n-k-1)}$$

- The restricted model now has no x variables, therefore the total variation explained by the restricted model must be zero, i.e. $R^2_r = 0$.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Under the null, the restricted model has only an intercept so

$$R^2_r = \frac{\sum(\overset{\hat{y}_i}{\bar{y}} - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Relationship between F-test and t-test

You might have found yourself asking "Could we have also used the F-test to test the individual significance of the parameters?"

- (Short & cautious answer is yes) In the case of testing a **single parameter restriction**, the F-test is actually the square of the t-test.

$$F_{1,n-k-1} = t^2_{n-k-1}$$

- We default to the t-test for a single parameter restriction because it is easier to compute.

Testing General Linear Restrictions

- Sometimes the restrictions implied by a theory are more complicated than just excluding some independent variables. For example,

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u$$

Suppose we want to test whether

1. the assessed house price is a rational valuation. If so, a 1% change in *assess* should be associated with a 1% change in *price*, hence $\beta_1 = 1$.
2. *lotsize*, *sqft*, and *bdrms* are all equally unimportant in determining the price of a house, once *assess* is controlled for.

Together, we have:

$$H_0 : \beta_1 = 1, \beta_2 = \beta_3 = \beta_4 = 0$$

Step 1: Estimate the Unrestricted Model

```r
(hed.mod <- lm(lprice ~ lassess + llotsize + lsqrft + bdrms,
               data = hprice1)) |> summary()
```

```
##
## Call:
## lm(formula = lprice ~ lassess + llotsize + lsqrft + bdrms, data = hprice
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5334 -0.0633  0.0069  0.0784  0.6083
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.26374    0.56966    0.46     0.64
## lassess      1.04307    0.15145    6.89    1e-09 ***
## llotsize     0.00744    0.03856    0.19     0.85
## lsqrft      -0.10324    0.13843   -0.75     0.46
## bdrms        0.03384    0.02210    1.53     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.148 on 83 degrees of freedom
```

Step 2: Estimate the Restricted Model

Under the null hypothesis, we can rewrite the model as:

$$log(price) = \beta_0 + \log(assess) + u$$

We could further redefine the model as:

$$\boxed{log(price) - \log(assess)} = \beta_0 + u$$

CAUTION!!!

Note that the dependent variable has changed. Therefore, **we cannot use the R^2 approach to test the null hypothesis** since **we cannot compare the R^2 of two models with different dependent variables**.

```r
(rest.hed <- lm(lprice - lassess ~ 1, data = hprice1)) |> summary()
```

```
##
## Call:
## lm(formula = lprice - lassess ~ 1, data = hprice1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5159 -0.0830 -0.0044  0.0850  0.6055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.0848     0.0157   -5.41  5.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.147 on 87 degrees of freedom
```

Step 3: Compute the F-statistic

```
# F-statistic
SSR_r <- sum(resid(rest.hed)^2)
SSR_ur <- sum(resid(hed.mod)^2)
df_r <- rest.hed$df.residual
df_ur <- hed.mod$df.residual
q <- df_r - df_ur
Fstat <- (SSR_r - SSR_ur)/q / (SSR_ur/df_ur)

# Fcrit = F(4,83, 5%)
Fcrit <- qf(p = 0.05, df1 = q, df2 = df_ur, lower.tail = FALSE)
cat("F-stat: ", Fstat, "\n",
    "Critical value: ", Fcrit, "\n",
    "Reject H0: ", Fstat > Fcrit)
```

```
## F-stat:  0.668
##  Critical value:  2.48
##  Reject H0:  FALSE
```

Conclusions?

```r
# F-statistic
car::linearHypothesis(hed.mod, c("lassess = 1", "llotsize = 0'
                                  "lsqrft = 0", "bdrms = 0"))
```

```
##
## Linear hypothesis test:
## lassess = 1
## llotsize = 0
## lsqrft = 0
## bdrms = 0
##
## Model 1: restricted model
## Model 2: lprice ~ lassess + llotsize + lsqrft + bdrms
##
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     87 1.88
## 2     83 1.82  4    0.0586 0.67   0.62
```

Economic vs Statistical Significance

1. Statistical significance
   - The F-test, t-test, and p-value are all statistical tests.
2. Economic significance
   - The economic significance of a parameter is the size (and sign) of the effect of the x on the dependent variable.
   - Addresses economic an policy relevance of the parameter.

Economic vs Statistical Significance

- A statistically significant parameter is not necessarily economically significant.
  - For example, a parameter may be statistically significant but have a very small effect on the dependent variable.
  - A common mistake is to overemphasize the statistical significance of a parameter without considering its economic significance.
- A statistically insignificant parameter is not necessarily economically insignificant.
  - For example, a parameter may be statistically insignificant but have a large effect on the dependent variable.
  - A common mistake is to discard an economically important variable because it is statistically insignificant.

Thought

Returning to the 2-year vs. 4-year college example, could you use the `car` package to test the hypothesis that returns to education are the same for both types of colleges?

```
# Using the car package
(jc.col <- car::linearHypothesis(two.year, c("jc = univ")))
```

```
##
## Linear hypothesis test:
## jc - univ = 0
##
## Model 1: restricted model
## Model 2: lwage ~ jc + univ + exper
##
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1   6760 1251
## 2   6759 1251  1     0.399 2.15   0.14
```

```
jc.col$F[2] |> sqrt() #t-stat equivalent
```

```
## [1] 1.47
```