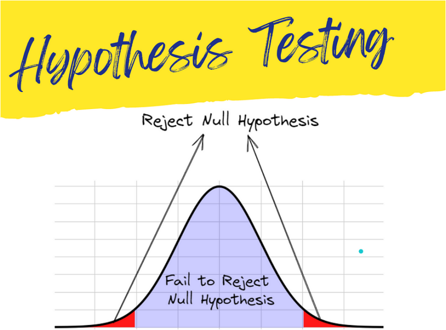
Fundamentals of Econometrics

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:

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

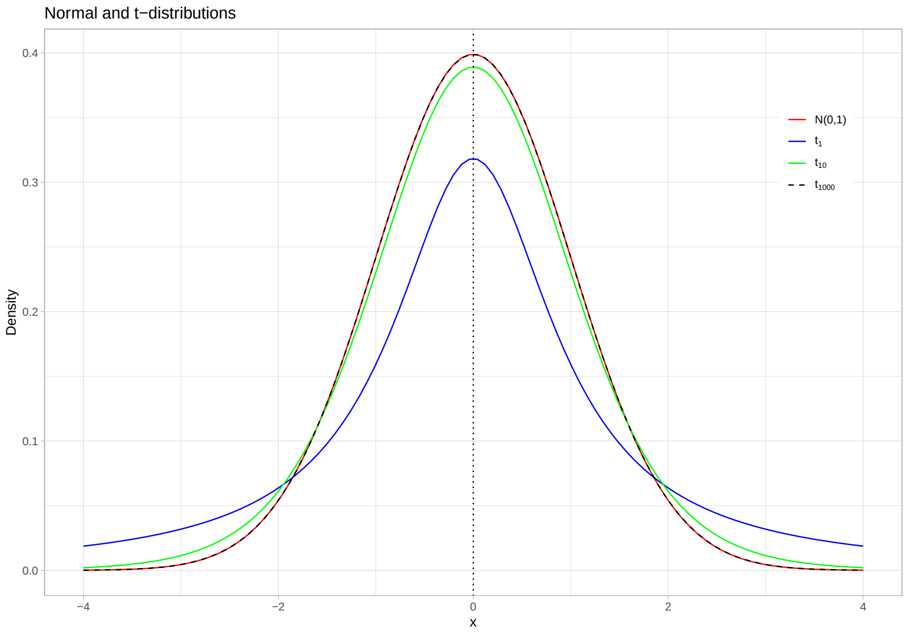
By the second property of the Normal distribution:

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

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}

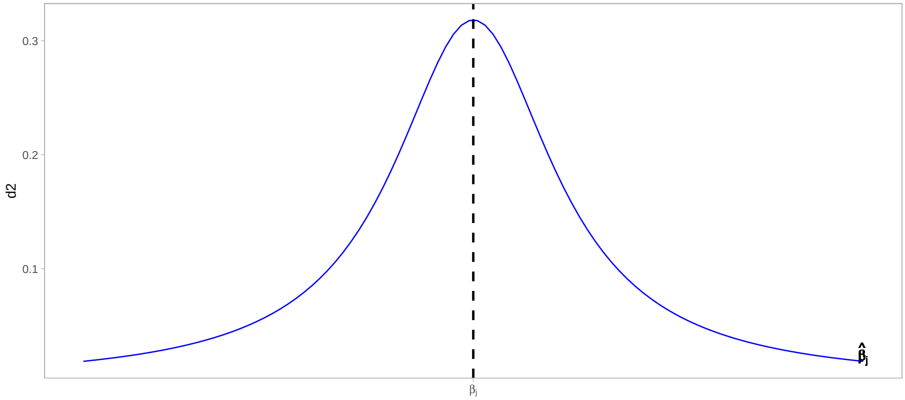
Section 2

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

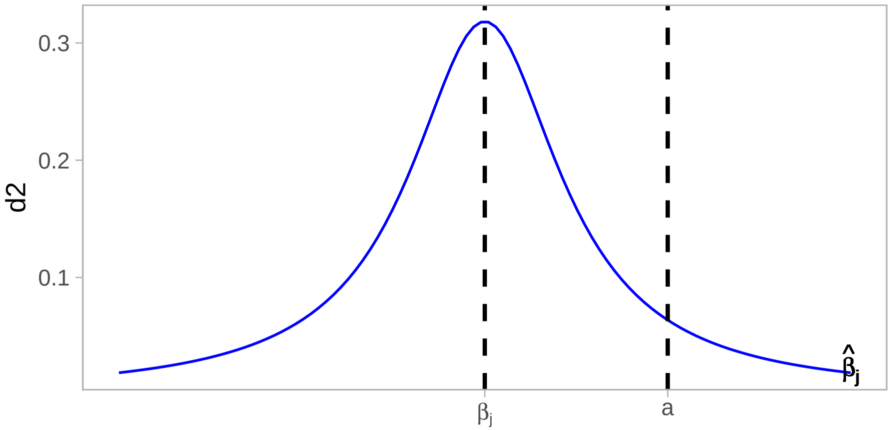
Our test statistics under the null hypothesis is:

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

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.



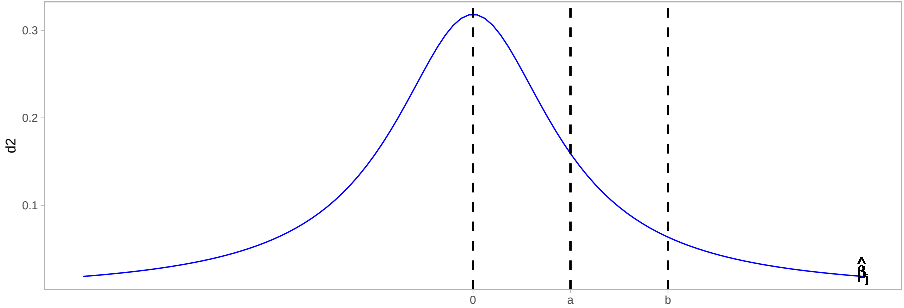
Why is the distribution of $\hat{\beta}_j$ important?

- 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:



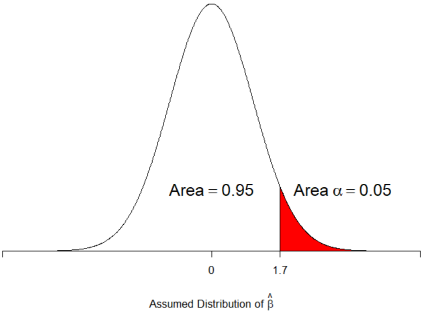
- At point `b`, we are more likely to reject than at point `a`.
- The basic question is "how do we know whether point `a` or point `b` is large enough to reject the null hypothesis?"

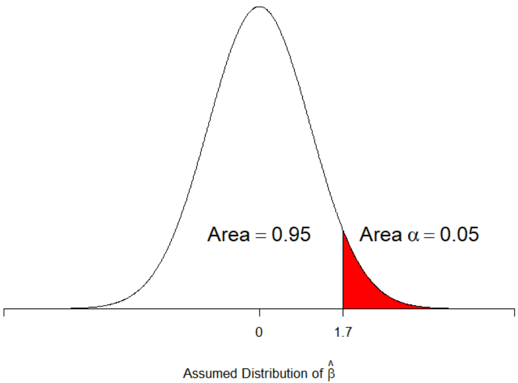
Critical Value

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
```

Finding probability values in R

```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
```

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?

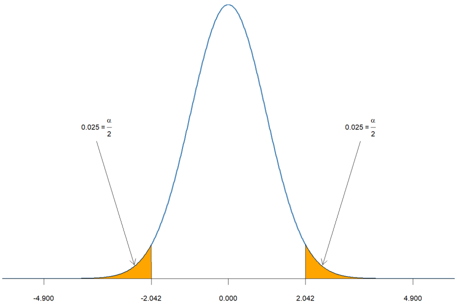
- Level of significance: α.
- dof, $n-k-1$:(2652)
- $t_{n-k-1,\alpha}$.

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?

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

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

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

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
```