Chapter2: Review of Simple Linear Regression

CZ

Fall, 2014

Review of Simple Linear Regression

A **simple linear** regression model is:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

Note:

- y is a random variable;
- In this course, x is always considered as non-random: measured without any error;
- β_0 and β_1 are unknown regression coefficients (fixed constants: frequentist); so randomness of y comes from ?
- Goal: estimate β_0 and β_1 based on a sample of observations (x_i, y_i), $i = 1, 2, \ldots, n$.

Assumptions

Suppose we observe n pairs of values: $(x_i, y_i), i = 1, \ldots, n$. For the ith observation, we have

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Important assumptions (Gauss-Markov assumptions) about the error term:

i $E(\epsilon_i) = 0$

ii $Var(\epsilon_i) = \sigma^2$

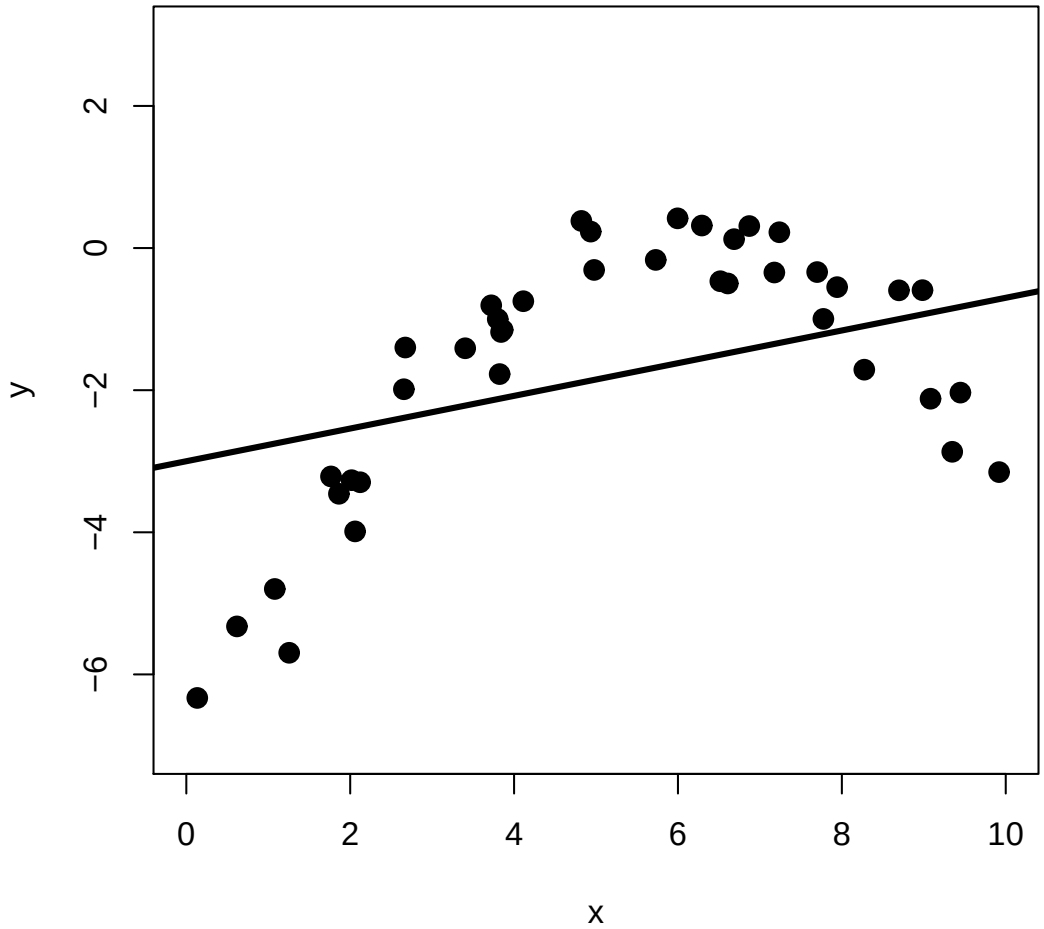
iii $\epsilon_1, \ldots, \epsilon_n$ are statistically independent

iv $\epsilon_i \sim N(0, \sigma^2)$.

Note: These four assumptions are often summarized as: $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$.
Question: which assumption is stronger?

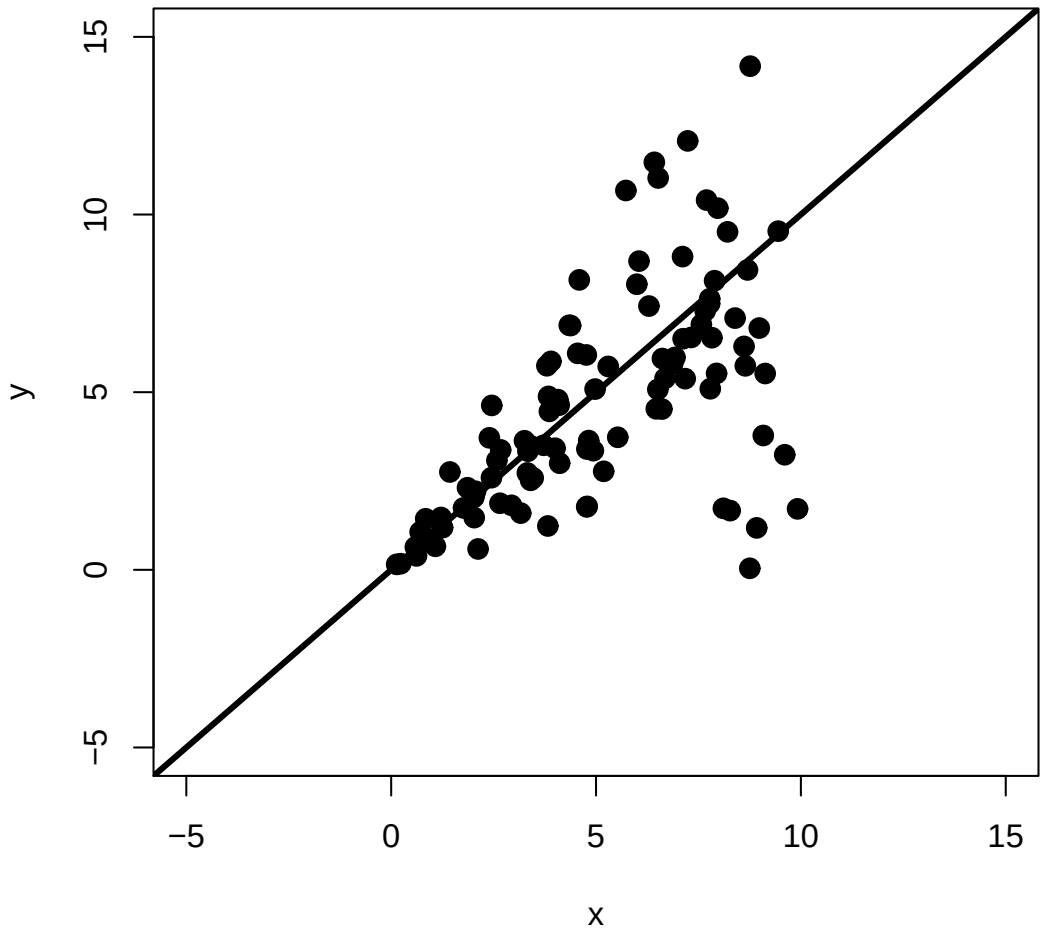
Assumptions

$E(\epsilon_i) = 0$: to ensure model is appropriate



Assumptions

$Var(\epsilon_i) = \sigma^2$: to ensure variance is consistent



Regression Coefficients

The assumptions imply that the mean value of y is a linear function of x:

$$\mu_i = E(y_i) = \beta_0 + \beta_1 x_i.$$

Interpret β_0 and β_1:

- β_1: **slope, of primary interest**
  $\beta_1 = E(y|x = a+1) - E(y|x = a)$: the average amount of change (increase/ decrease) in response when the value of the predictor increases by 1 unit.
- β_0: **intercept**
  $\beta_0 = E(y|x = 0)$: the average value of y when $x = 0$.

Estimation

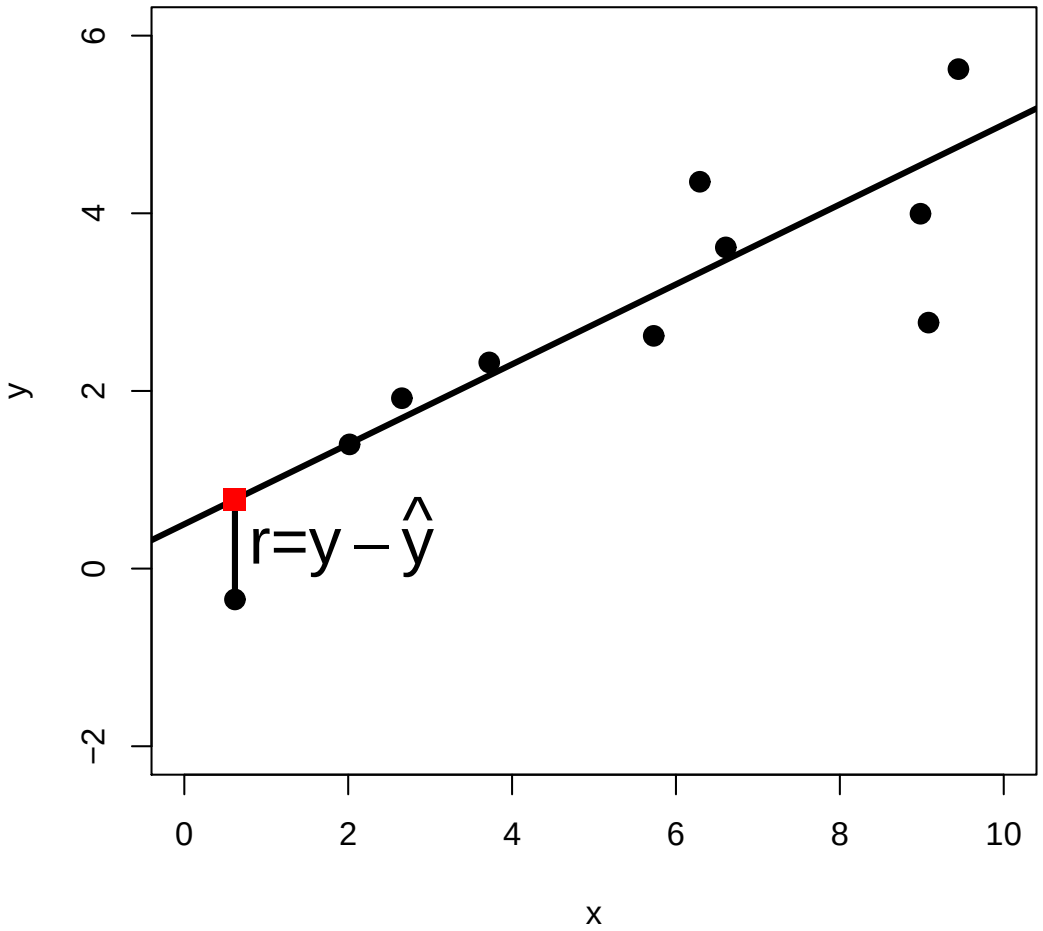
Denote the estimators of β_0, β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$. The fitted model is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i; \quad i = 1, ..., n.$$

How to decide the best values $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the sample?

$$\begin{aligned} \text{Predicted/Fitted values} : \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \text{Observed errors (residuals)} : r_i &= y_i - \hat{y}_i \\ \text{Sum of squared errors} : SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \end{aligned}$$

Estimation



Least Square Estimators (LSEs)

Denote

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2}$$

Differentiate (2) with respect to β_0 and β_1, we get:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i) \tag{3}$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) \tag{4}$$

Set the partial derivatives (3) and (4) equal to zero, using $\hat{\beta}_0$ and $\hat{\beta}_1$ to denote the estimates, we obtain:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Inner steps

Consequence of LS fitting

For least square estimation, we have:

1. $\sum r_i = 0$ (from (3))
2. $\sum r_i x_i = 0$ (from (4))
3. $\sum r_i \hat{y}_i = 0$ (why?)
4. The point $(\bar{x}, \bar{y})$ is always on the fitted regression line (why?)

An Example about the Simple Regression Model

Example: The simple regression equation relating height (y) and distance between fingertips (DF, x) is

$$\text{average height} = \beta_0 + \beta_1 \cdot \text{DF}.$$

Suppose the sample is:

DF (cm)	156	176	167	155	180	178	145	177	189
Height (cm)	153	171	163	150	180	188	142	182	180
DF (cm)	165	176	178	182	158	163	171	150	188
Height (cm)	160	173	176	185	158	165	169	154	186

An Example about the Simple Regression Model

Calculate $\bar{x} = 169.668$, $\bar{y} = 168.611$, $S_{xx} = 2850$, $S_{xy} = 2856.667$. Then the sample intercept is $\hat{\beta}_0 = -1.452$ and the sample slope is $\hat{\beta}_1 = 1.002$. Interpret the parameters:

- $\hat{\beta}_0$: the estimated average height at DF=0 is -1.452 cm.
- $\hat{\beta}_1$: For one unit increase in DF, the estimated average height increases by 1.002 cm.

$$\widehat{\text{height}} = -1.452 + 1.002 \cdot \text{DF}.$$

Properties of Least Squares Estimators

We have following properties of LSEs

- Property 1. LSEs are unbiased:

$$E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_0) = \beta_0$$

- Property 2. The theoretical variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \\ Var(\hat{\beta}_0) &= \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}] \end{aligned}$$

- Property 3. $\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2\bar{x}}{\sum(x_i - \bar{x})^2}$

Property 4. Under the normality assumption (iv), we have:

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2})$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}])$$

Proof

Proof

Proof

Point Estimator of σ^2

Recall: $\sigma^2 = Var(\epsilon_i)$. We could say that $r_i = y_i - \hat{y}_i$ (somehow) estimates the unobservable ϵ_i.
The idea is then to use the sample variance of $r_1, r_2, \ldots, r_n$ to estimate the unknown σ^2:

$$\frac{1}{n-1}\sum(r_i - \bar{r})^2 = \frac{1}{n-1}\sum r_i^2 \tag{5}$$

because $\bar{r}$=0. However, this is biased:

$$E(\frac{1}{n-1}\sum r_i^2) \neq \sigma^2$$

Cont.

Nevertheless, if we define

$$s^2 = \frac{\sum r_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

we have,

$$E\{s^2\} = \sigma^2$$

s^2 is an unbiased estimator of σ^2 and is called mean square error (**MSE**).
Note: The denominator $n-2$ only applies to simple regression (to estimate β_1 and β_0). The general rule is that the denominator is $n-p$, where p is the number of parameters in the regression equation.

Proof

Proof

Statistical Inference

Inference: theoretical properties about estimators ($\hat{\beta}_1$, etc.) or related concepts

Basically, two main statistical inference tools:

- Confidence interval
- Hypothesis testing

They are almost equivalent, sort of.

Distribution of $\hat{\beta}_1$, $\hat{\beta}_0$

Under the normality assumption $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$, we have $\hat{\beta}_1$ and $\hat{\beta}_0$ are normally distributed and:

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2})$$

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}])$$

Estimated Variance

Recall: s^2 (MSE) is an unbiased estimator of σ^2. Replace the unknown parameter σ^2 with s^2, we obtain:

$$se(\hat{\beta}_1) = \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}$$

$$se(\hat{\beta}_0) = \sqrt{s^2 \cdot [\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}]}$$

Standard error VS. Standard deviation
`http://www-ist.massey.ac.nz/dstirlin/CAST/CAST/HseMean/seMean7.html`

Distribution of $\hat{\beta}_1$, $\hat{\beta}_0$

Since $\hat{\beta}_1$ and $\hat{\beta}_0$ are normally distributed, we know the standardized statistic

$$\frac{\hat{\beta}_1 - \beta_1}{sd(\hat{\beta}_1)} \sim N(0,1)$$

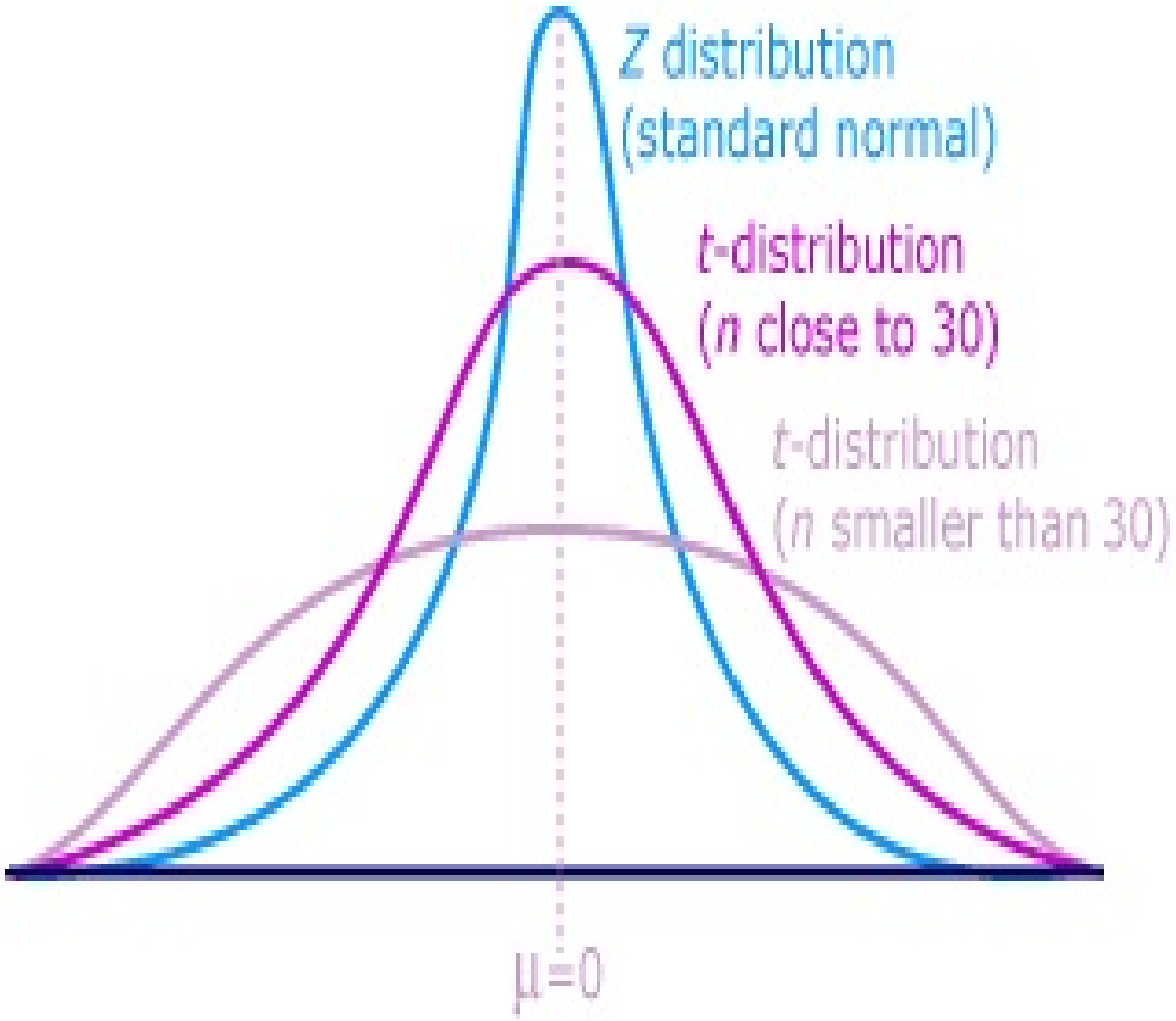
$$\frac{\hat{\beta}_0 - \beta_0}{sd(\hat{\beta}_0)} \sim N(0,1)$$

However if we replace σ^2 by its unbiased estimate s^2, we have:

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2}$$

$$\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim t_{n-2}$$

t-distribution and Z-distribution



Confidence Intervals

- A **confidence interval (CI)** is an "interval estimate" of the population parameter, i.e., an interval of values that is likely to include the unknown value of the population parameter.
- The **confidence level** is the probability that the random interval "captures" the true value of the population parameter, often denoted by $100(1-\alpha)\%$. As an example, a 95% confidence interval means: among 100 random samples, 95 of them are likely to "capture" the population value.
- The higher the confidence level is, the wider the confidence interval should be.

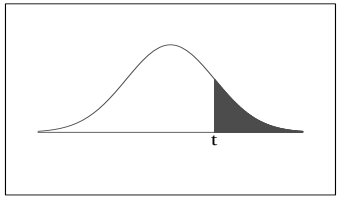
Confidence Interval for Slope β_1

A $100(1-\alpha)\%$ confidence interval for the unknown slope β_1 can be computed as

$$\begin{aligned} (\text{point estimate} &\pm \text{Multiplier } t^\star \times \text{se of the point estimate}) \\ (\hat{\beta}_1 &\pm t_{n-2,\alpha/2} \times se(\hat{\beta}_1)). \end{aligned}$$

where $t_{n-2,\alpha/2}$ is the **upper** $100(\alpha/2)th$ percentile for t_{n-2}.

t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
∞	1.282	1.645	1.960	2.326	2.576

Hypothesis Testing for slope β_1

The slope directly tells us about the link between mean y and x.

- If $\beta_1 \neq 0$, the variables y and x are **linearly** related.
- If $\beta_1 = 0$, there is no **linear** relationship because mean y does not change when the value of x is changed.

Example: The simple regression equation relating height (y) and DF (x) is

$$\text{average height} = \beta_0 + \beta_1 \cdot \text{DF}.$$
$$\Rightarrow \quad \widehat{\text{height}} = -1.452 + 1.002 \cdot \text{DF}.$$

But is this linear relationship "statistically significant"?

Hypothesis Testing for slope β_1

- Step 1: the null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_1^*, \ H_a : \beta_1 \neq \beta_1^*$$

- Step 2: Construct a test statistic:
  - A test statistic should not contain the unknown parameter.
  - The test statistic is a random variable.
  - The distribution of the test statistic should be known.

  Start with the distributions of $\hat{\beta}_1$:

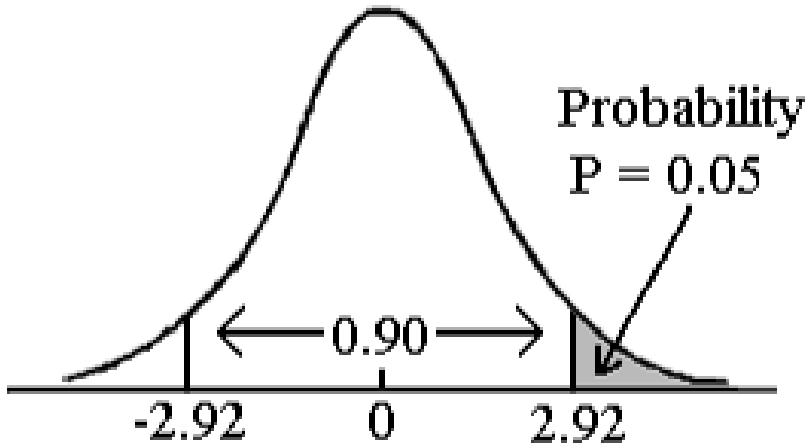
$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2}$$

  Hence, under $H_0 : \beta_1 = \beta_1^*$, T below can serve as a test statistic:

$$T = \frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)} \sim t_{n-2}$$

Hypothesis Testing for slope β_1

- Step 3: Make conclusion:
  Under H_0, the density curve of the constructed T variable is bell-shaped and symmetric at 0.



The area under the t-curve gives us the probability of the variable taking values in a certain interval.

Hypothesis Testing for slope β_1

(1) **Critical value approach:**

- If H_0 is true, that is, the random variable $T \sim t_{n-2}$, then the sample value of T, denoted by T_0 (T_0 is the value of T given the specific sample), should have a large chance to fall in the middle area, that is, T_0 should be close to 0.

- Therefore, if $|T_0|$ is "too large", H_0 is likely to be wrong. In practice, we use $|T_0| > t_{n-2,\alpha/2}$ to indicate "too large" and that we should reject H_0, e.g. $\alpha = 0.05$.
  α is called the significance level.

Hypothesis Testing for slope β_1

(2) **p-value approach:**

- If H_0 is true, the sample value T_0 has a large chance to fall in the middle area, i.e. $|T_0|$ is small.
- Then the probability that random variable T is more extreme than T_0 is large, i.e. p-value$= P(|T| > |T_0|)$ should be large.
- Therefore, if p-value is "very small", H_0 is likely to be wrong. In practice, we use p-value$< \alpha$ to indicate "very small", e.g. $\alpha = 0.05$.

Hypothesis Testing for slope β_1

In summary,

(1) **Critical value approach:**

- if $|T_0| > t_{n-2,\alpha/2}$, we reject H_0 and conclude $\beta_1 \neq \beta_1^*$
- if $|T_0| \leq t_{n-2,\alpha/2}$, we don't have enough evidence to reject H_0 and conclude $\beta_1 = \beta_1^*$

(2) **p-value approach:**

- if p-value $< \alpha$, we reject H_0 and conclude $\beta_1 \neq \beta_1^*$
- if p-value $\geq \alpha$, we don't have enough evidence to reject H_0 and conclude $\beta_1 = \beta_1^*$

The significance level is usually $\alpha = 0.05$.

Hypothesis Testing v.s. Confidence Interval

They are equivalent in some sense. Consider the hypothesis testing at the significance level of α:

$$H_0 : \beta_1 = 0, \ H_a : \beta_1 \neq 0$$

Reject the null hypothesis if $\mathbf{0}$ is not included in the $100(1-\alpha)\%$ confidence interval for the slope.

Sampling Distribution of $\hat{y}_0$

After β_0, β_1 are estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$, the fitted value given an x_0 is: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

- $E(\hat{y}_0) = E(y_0)$
  - $E(\hat{y}_0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_0 = \beta_0 + \beta_1 x_0$
  - $E(y_0) =$
- $Var(\hat{y}_0) = \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]$ (Verify yourself)
- Under normality assumption,

$$\hat{y}_0 \sim N(E(y_0), \sigma^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}])$$

- $\frac{\hat{y}_0 - E(y_0)}{se(\hat{y}_0)} \sim t_{n-2}$ where $se(\hat{y}_0) = \sqrt{s^2[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}]}$.

Confidence Interval for $E(y_0)$

- The confidence interval for $E(y_0)$ is given by

$$\begin{aligned} (\text{point estimate} &\pm \text{Multiplier } t^\star \times \text{se of the point estimate}) \\ (\widehat{y_0} &\pm t_{n-2,\alpha/2} \times se(\hat{y}_0)) \end{aligned}$$

where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, and $se(\hat{y}_0) = \sqrt{s^2 \cdot (\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2})}$.

Prediction Interval for y_p

- A **prediction interval for a new y (PI)** is an interval estimate for a new individual observation y_p (random variable) given a new x_p.
- The new observation on y_p to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based.
- The new observation can be written as

$$y_p = \beta_0 + \beta_1 x_p + \epsilon_p.$$

  where ϵ_p is a future unknown random error.
- After β_0, β_1 are estimated by $\hat{\beta}_0$ and $\hat{\beta}_1$, the point estimator for new y_p given a new x_p is:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 x_p.$$

Properties of $y_p - \hat{y}_p$

- $E(y_p - \hat{y}_p) = 0$ (Similar as y_0 and $\hat{y}_0$; compare to previous slides)
- $Var(y_p - \hat{y}_p) = \sigma^2 \cdot (1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2})$ (WHY?? Think it in terms of: where does randomness come from?)
- Under normality assumption,

$$y_p - \hat{y}_p \sim N(0, \sigma^2 \cdot (1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2})$$

- $\frac{y_p - \hat{y}_p}{se(y_p - \hat{y}_p)} \sim t_{n-2}$ where $se(y_p - \hat{y}_p) = \sqrt{s^2 \cdot (1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2})}$

CI for $E(Y)$ and PI for y: Comparison

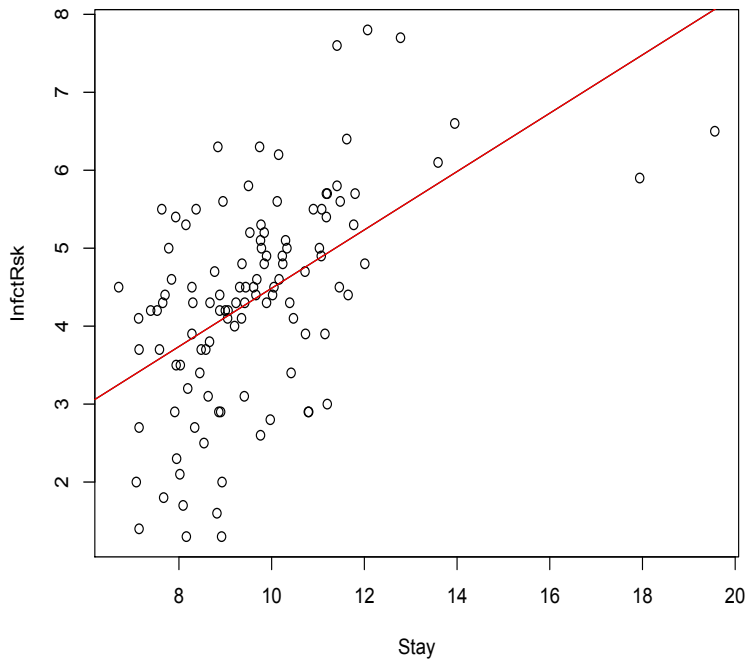
- A **prediction interval for** y_p at x_p is calculated as

$$\hat{y}_p \pm t_{n-2,\alpha/2}\sqrt{s^2 \cdot (1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2})}$$

Therefore, P.I. is always **wider** than the corresponding C.I.

Example

Example: Consider a sample of $n = 113$ hospitals in the east and north central U.S. The response variable for this dataset is y =infection risk (% of patients who get an infection), and the predictor variable is x =average length of stay (days).



R Output

```
Call:
lm(formula = InfctRsk ~ Stay, data = senic)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7823 -0.7039  0.1281  0.6767  2.5859

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74430    0.55386   1.344    0.182         Row (1)
Stay         0.37422    0.05632   6.645 1.18e-09 ***     Row (2)
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.139 on 111 degrees of freedom
Multiple R-squared: 0.2846, Adjusted R-squared: 0.2781
F-statistic: 44.15 on 1 and 111 DF,  p-value: 1.177e-09
```

Hypothesis Test for Intercept: Example

Row(1) in the output gives information used to make inference about the **intercept**. The null and alternative hypotheses for a hypotheses test about the intercept are

$$H_0 : \beta_0 = 0, \ H_1 : \beta_0 \neq 0$$

- The test statistic based on the given is

$$T_0 = \hat{\beta}_0 / se(\hat{\beta}_0) = 0.74430/0.55386 = 1.344,$$

  and the cut value

$$t_{n-2,\alpha/2} = t_{113-2,0.05/2} = 1.98,$$

  hence $|T_0| < t_{n-2,\alpha/2}$, indicating that we don't have enough evidence to reject H_0 and the intercept is **not** significant.
- p-value$>$ 0.05, given the same conclusion.

Hypothesis Test for Slope: Example

Row(2) in the output on the previous page gives information used to make inferences about the **slope**. The null and alternative hypotheses for a hypotheses test about the slope are

$$H_0 : \beta_1 = 0, \ H_1 : \beta_1 \neq 0$$

- The test statistic based on the given is

$$T_0 = \hat{\beta}_1 / se(\hat{\beta}_1) = 0.37422/0.05632 = 6.645,$$

  and the cut value

$$t_{n-2,\alpha/2} = t_{113-2,0.05/2} = 1.98,$$

  hence $|T_0| > t_{n-2,\alpha/2}$, indicating that we should reject H_0 and the linear relation is significant.
- p-value$\approx$ 0, given the same conclusion.

CI for $E(y)$ and PI for y: Example

```
> predict(model, newdata=data.frame(Stay=10),
interval="confidence",se.fit=TRUE, level=0.95)
$fit
       fit      lwr      upr
1 4.486472 4.270502 4.702443
$se.fit
[1] 0.1089897
$df
[1] 111

> predict(model, newdata=data.frame(Stay=10),
interval="prediction",se.fit=TRUE, level=0.95)
$fit
       fit      lwr      upr
1 4.486472 2.218596 6.754349
$se.fit
[1] 0.1089897
$df
[1] 111
```

CI for $E(y)$ and PI for y: Example

Interpretation:

- "95% CI (lwr, upr)":
  With 95% confidence we can estimate that in hospitals where the average length of stay is 10 days, the **mean infection risk** is between 4.271 and 4.702.
- "95% PI (lwr, upr)":
  For a new hospital with average length of stay is 10 days, we have 95% confidence to predict that its **infection risk** is between 2.219 and 6.754.
- "fit":
  is calculated as $\hat{y} = 0.7443 + 0.3742 \times 10 = 4.486$.
- "se.fit":
  is the standard error of $\hat{y}$; it measures the accuracy of $\hat{y}$ as an estimate of $E(y)$.

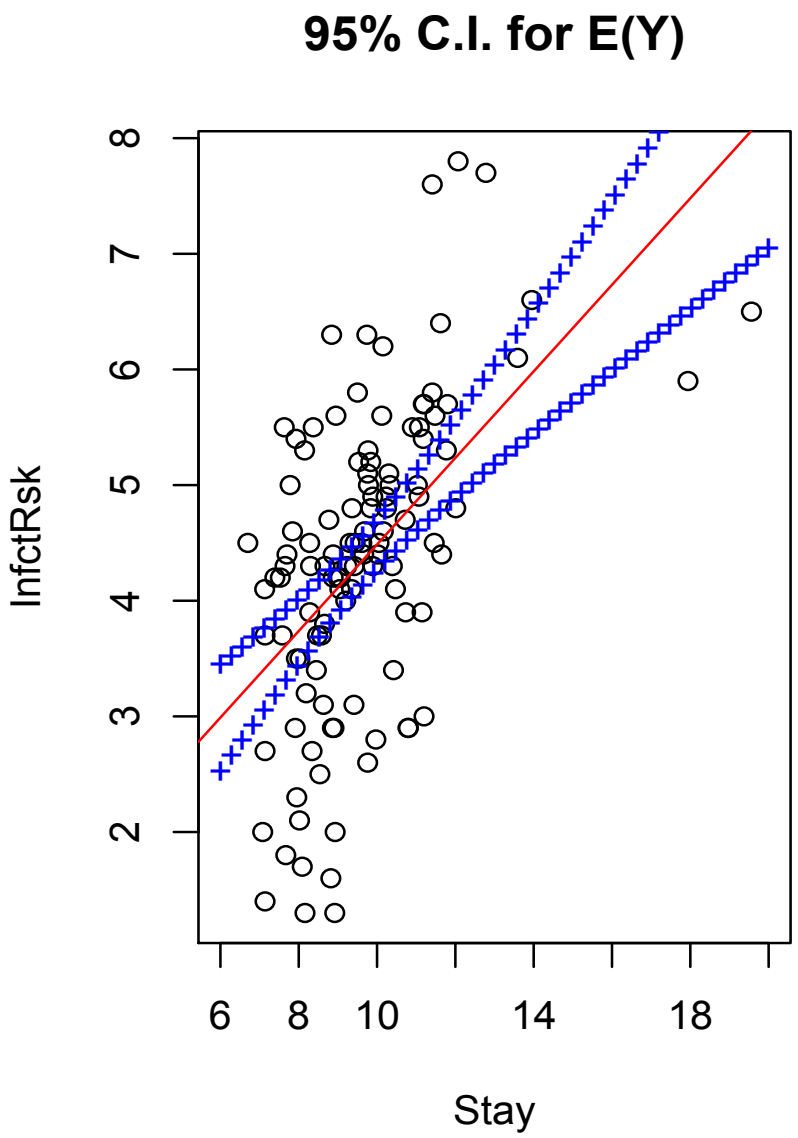
CI for $E(Y)$ and PI for y: Example

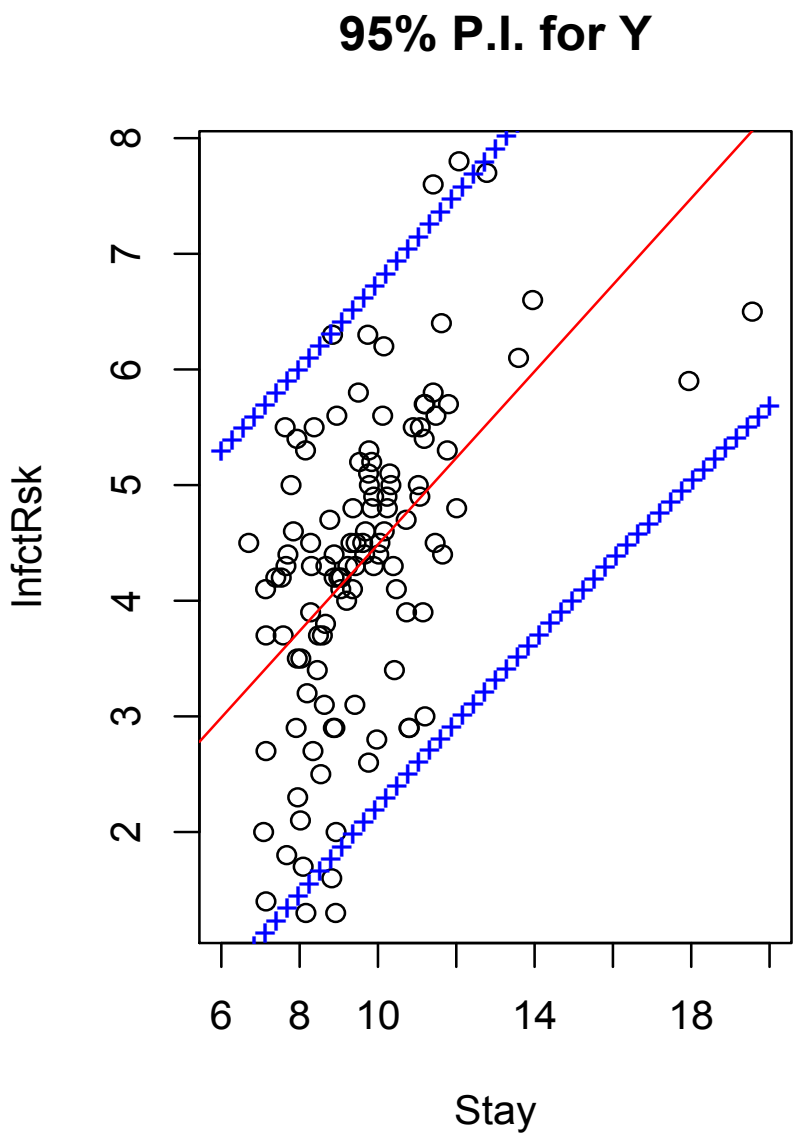
"By hand" calculation:

$d.f. = n - 2 = 111, t^\star = t_{n-2,0.025} = 1.98, s = 1.139$, so

$$\begin{aligned} \text{CI for } E(y) &: (4.486 \pm (1.98 \times 0.109)) = (4.270, 4.702); \\ \text{PI for new } y &: (4.486 \pm (1.98 \times \sqrt{1.139^2 + 0.109^2})) \\ &= (2.220, 6.752). \end{aligned}$$

CI for $E(y)$ and PI for y: Example





Analysis of Variance (ANOVA)

- An **analysis of variance (ANOVA)** table for regression displays quantities that measure how much of the variability in the y-variable is explained and is not explained by the regression relationship with the x-variable(s).
- It provides an alternative way to test $H_0 : \beta_1 = 0$

Analysis of Variance (ANOVA): Main Quantities

$$SST = SSE + SSR$$

- Total Sum of Squares (SST):
  $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is a measure of the variation in observed y values, calculated independently of the x values. $df(SST) = n - 1$.
- Sums of Squared Errors or Residual Sum of Squares (SSE):
  $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is a measure of variation in y NOT explained by the regression. $df(SSE) = n - p - 1$, where p is **number of predictors** in the model. For simple regression, $p = 1$.
- Regression Sum of Squares (SSR): $SSR = SST - SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is a measure of the total variation in y that can explained by the regression model. $df(SSR) = p$.

Analysis of Variance (ANOVA): Other Quantities

- Mean Squared Error (MSE):
  $MSE = SSE/(n-p-1)$, can serve as the estimate of σ^2, the variance of ϵ.
- Mean Square for the Regression (MSR):
  $MSR = SSR/p$, is used to construct F test statistic (introduced later) along with MSE.

Analysis of Variance (ANOVA): Table

Source	DF	SS	$MS = SS/DF$	F
Regression	p	SSR	MSR	MSR/MSE
Error	$n - p - 1$	SSE	MSE	
Total	$n - 1$	SST		

A little exercise

Verify that

- $\text{SST} = S_{yy}$
- $\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$
- $\text{SSR} = \frac{S_{xy}^2}{S_{xx}}$

F-test

For simple linear regression ($p = 1$), if $H_0 : \beta_1 = 0$ is true, we have:

1. $\frac{SST}{\sigma^2} \sim \chi^2(n-1)$
2. $\frac{SSR}{\sigma^2} \sim \chi^2(1)$
3. SSE is independent of SSR and $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$
4. SSR independent with SSE under Assumptions (iii) and (iv)

The F statistic is calculated as:

$$F = \frac{MSR}{MSE}$$

Under $H_0 : \beta_1 = 0$, we have: $F \sim F(1, n-2)$.

Proof

$\frac{SST}{\sigma^2} \sim \chi^2(n-1)$

Proof

$\frac{SSR}{\sigma^2} \sim \chi^2(1)$

F-test

- **Critical value approach:**
  H_0 is rejected if the calculated statistic, F_0, is such that:

$$F_0 > F_\alpha(1, n-2),$$

  where $F_\alpha(1, n-2)$ is the α **upper** percentile (or equivalently, $1-\alpha$ percentile) for $F(1, n-2)$.
- **p-value approach:**
  p-value$= P(F > F_0)$ where $F \sim F(1, n-2)$
  If p-value$< \alpha$, reject H_0.

t-test vs. F-test

NOTE:

- For **simple linear regression**, t-test and F-test for β_1 are equivalent.
- However, for **multiple regression**, while t-test is used to test the significance of each β coefficient, F-test is used to test the following hypothesis:

$$
\begin{aligned}
H_0 &: \beta_1 = \beta_2 = ... = \beta_p = 0 \\
H_a &: \text{at least one of the } \beta_i \neq 0, \text{ for } i = 1, ..., p.
\end{aligned}
$$

Coefficient of Determination R^2

$$R^2 = \frac{SSR}{SST}$$

- $0 \leq R^2 \leq 1$
- R: same as the correlation coefficient in STAT 230 **if simple linear regression**
- It is the proportion of the total variation in y that is explained by the regression model.