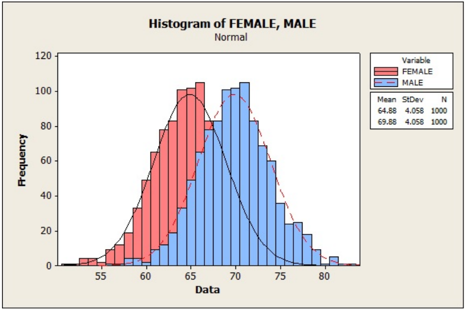
About Normal Distribution

REM: variance properties
Var(cX) = c^2 Var(X)
Var(c+X) = Var(X)



This is a simulation to demonstrate:
If **W** follows normal distribution with mean $\boldsymbol{\mu}$, variance $\boldsymbol{\sigma}^2$
Then **W+2** follows normal distribution with mean $\boldsymbol{\mu}+2$, variance $\boldsymbol{\sigma}^2$
Or simply $W \sim N(\mu, \sigma^2) \Rightarrow W + 2 \sim N(\mu + 2, \sigma^2)$.

About Normal Distribution

Properties of Normal Distribution

- If $W \sim N(\mu, \sigma^2)$, then $W + a \sim N(\mu + a, \sigma^2)$, where a is a constant.
- If $W \sim N(\mu, \sigma^2)$, then $\frac{W-\mu}{\sigma} \sim N(0, 1)$ standardized by z score called a standard normal distribution or Z-distribution.
- *If $W \sim N(\mu, \sigma^2)$, then $a \times W \sim N(a\mu, a^2\sigma^2)$, where a is a constant.*
- *If $W_1 \sim N(\mu_1, \sigma_1^2)$, $W_2 \sim N(\mu_2, \sigma_2^2)$, and W_1 and W_2 are independent, then $W_1 + W_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

About Normal Distribution

Recall for simple linear regression,

$$Y = \beta_0 + \beta_1 X + \epsilon$$

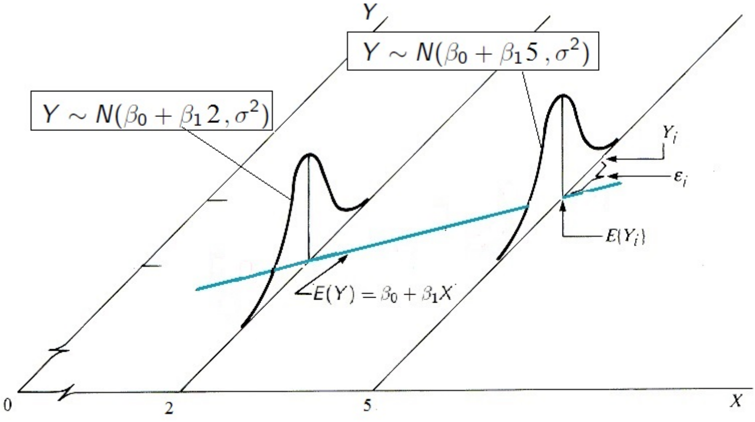
We usually assume

- ϵ follows normal distribution with mean 0 and variance σ^2 (usually unknown).
- β_0, β_1 are population parameters(fixed constant).
- X are known constant.

so Y follows normal distribution with mean $\beta_0 + \beta_1 X$, variance σ^2, or simply

$$Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

Regression Demonstration



The distributions of Y are different at different X values.
e.g Galton experiment on size of mother pea and daughter pea.

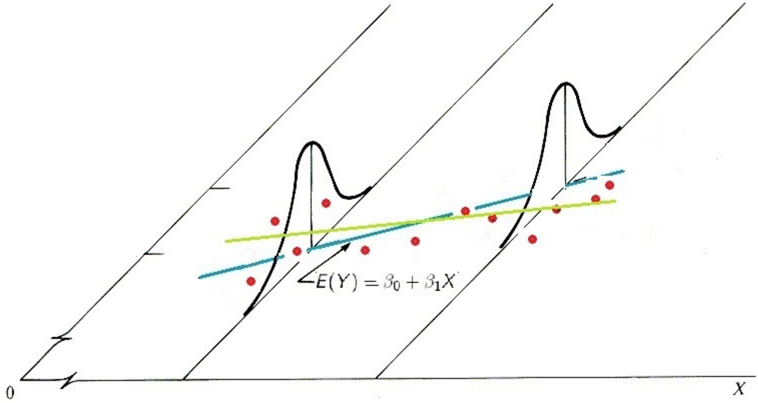
Sampling Distribution

- **Interest:**
  Is there really linear relationship between sizes of mother pea and sizes of daughter pea?
  Given a mother pea of a certain size, what are most possible sizes of daughter peas?
  How can we predict sizes of daughter peas based on a new size of mother pea?
  ...
- **Problem:** We don't know β_0 and β_1!
- **Solution:** We will use sample estimates b_0 and b_1 to estimate and make inference.

  Why this is true? How to make inference?
  Think about the experiment that Galton did.

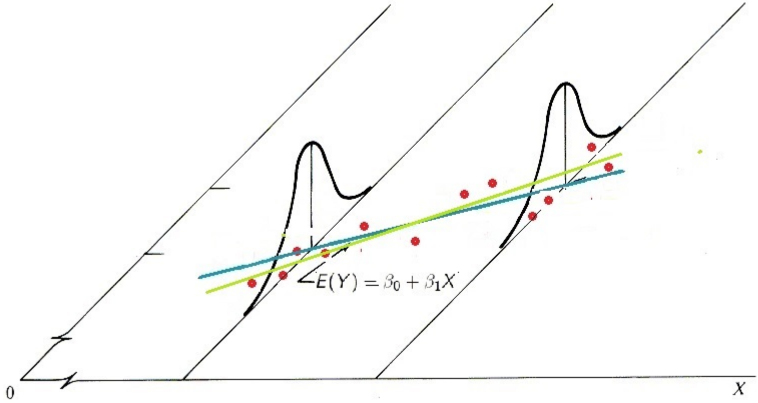
Sampling Distribution of b_0 and b_1

Friend 1 (mother pea size and daughter pea size)



Sampling Distribution of b_0 and b_1

Friend 2 (mother pea size and daughter pea size)



Sampling Distribution of b_0 and b_1

- Assume Galton has asked many friends to help.

$$
\begin{array}{lcll}
\text{friend 1 } \{(ms_1, ds_1^{(1)}), \cdots, (ms_n, ds_n^{(1)})\} & \rightarrow & b_0^{(1)} & b_1^{(1)} \\
\text{friend 2 } \{(ms_1, ds_1^{(2)}), \cdots, (ms_n, ds_n^{(2)})\} & \rightarrow & b_0^{(2)} & b_1^{(2)} \\
\cdots & \rightarrow & \cdots & \cdots \\
\text{friend n } \{(ms_1, ds_1^{(n)}), \cdots, (ms_n, ds_n^{(m)})\} & \rightarrow & b_0^{(m)} & b_1^{(m)} \\
\cdots & \rightarrow & \cdots & \cdots
\end{array}
$$

  ms:mother pea size, ds: daughter pea size
- The distribution of $\{b_0^{(1)}, \cdots, b_0^{(m)}, \cdots\}$ and $\{b_1^{(1)}, \cdots, b_1^{(m)}, \cdots\}$ are sampling distributions

Sampling Distribution of b_0 and b_1

b_1 and b_0 are normally distributed and:

$$b_1 \sim N(\beta_1, Var(b_0))$$

$$b_0 \sim N(\beta_0, Var(b_1))$$

(These can also be verified by mathematical proofs)

$$Var(b_0) = \sigma^2 \cdot [\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}], Var(b_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Distribution of b_1, b_0

Since b_1 and b_0 are normally distributed, we know the standardized statistic

$$\frac{b_1 - \beta_1}{\sqrt{Var(b_1)}} \sim N(0,1), \quad \frac{b_0 - \beta_0}{\sqrt{Var(b_0)}} \sim N(0,1)$$

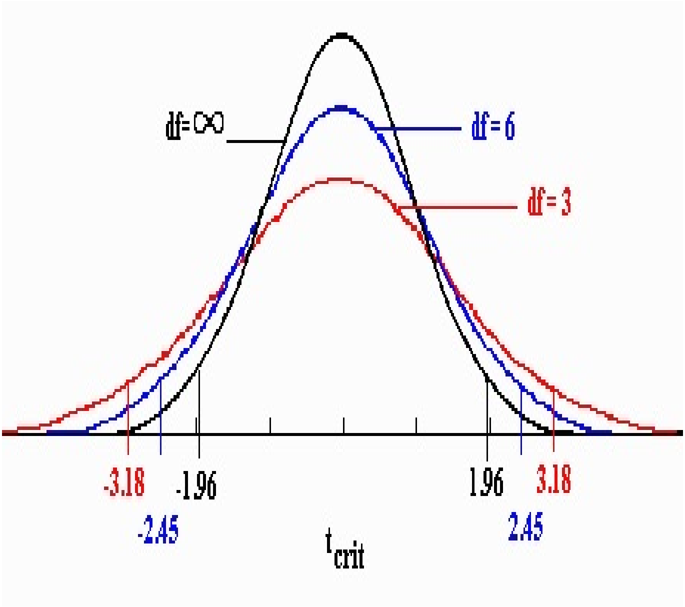
However if we replace Var by estimate* s^2 (usually reported in R), we have:

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$$

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t(n-2)$$

*Replace σ^2 by MSE

t-distribution and Z-distribution



Statistical Inference Tools

Basically, there are two main statistical inference tools:

- Hypothesis testing
- Confidence interval

They are equivalent in some sense.

Hypothesis Testing

Basic idea:
We are interested in whether the population parameter equals to a specific value or falls into a certain interval of possible values. Then, we can state two hypotheses, called null (H_0) and alternative (H_1 or H_a) hypotheses respectively, each of which contains some possible value(s) of the population parameter.

Hypothesis Testing

- Step 1: State the null and alternative hypotheses.
- Step 2: A test statistic is calculated using the sample.
- Step 3: Make conclusion–there are two completely equivalent strategies for making a decision:

Hypothesis Testing

(1) **Critical value approach:** We decide in favor of the alternative hypothesis when the value of the test statistic is more extreme than a critical value. The critical value is determined by the distribution of test statistic and the significance level. The significance level is usually $\alpha = 0.05$.

(2) **p-value approach:** This is used by all statistical software. We find the probability that the test statistic would be as extreme as is observed, if the null hypothesis were true. We decide in favor of the alternative hypothesis (over the null) when the p-value is less than the significance level. The significance level is usually set at $\alpha = 0.05$.

Hypothesis Testing for slope β_1

For instance, we usually want to know whether the slope of the simple regression model equals 0 or not, since the slope directly tells us about the link between mean y and x. When the true population slope β_1 does not equal 0, the variables y and x are linearly related. When the slope is 0, there is no linear relationship because mean y does not change when the value of x is changed.

Hypothesis Testing for slope β_1

- Step 1: the null and alternative hypotheses:

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

- 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 sampling distributions of b_1:

$$\frac{b_1 - \beta_1}{s(b_1)} \sim t(n-2)$$

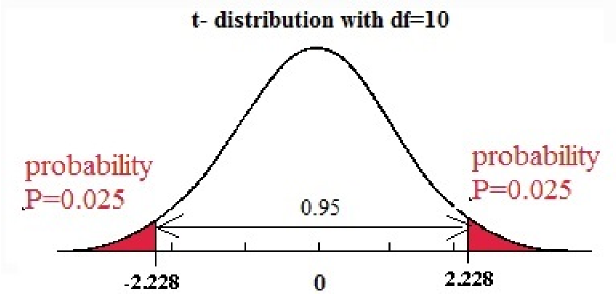
  Hence, T below can serve as a test statistic:

$$T = \frac{b_1}{s(b_1)} \sim t(n-2) \;\; \text{under } H_0 : \beta_1 = 0.$$

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.

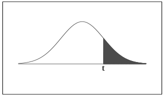
Hypothesis Testing for slope β_1

(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 $|T_0| > t_{\alpha/2}(n-2)$ to indicate "too large" and that we should reject H_0, e.g. $\alpha = 0.05$.

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

(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 intercept β_0

- Step 1: the null and alternative hypotheses:

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

- Step 2: Construct a test statistic:

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t(n-2)$$

Hence, T below can serve as a test statistic:

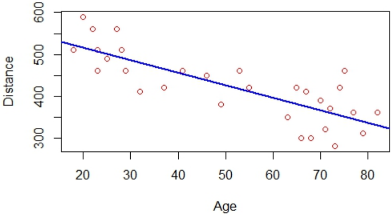
$$T = \frac{b_0}{s(b_0)} \sim t(n-2) \;\; \text{under } H_0 : \beta_0 = 0.$$

- Step 3: Make conclusion:

Same as for the slope.

Hypothesis Testing: Example

Example: $n = 30$ observations on driver age and the maximum distance (feet) at which individuals can read a highway sign:




```
Call:
lm(formula = Distance ~ Age, data = sign)

Residuals:
    Min      1Q  Median      3Q     Max 
-78.231 -41.710   7.646  33.552 108.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
Age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07
```

Hypothesis Test for Intercept: Example

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

  $$T_0 = b_0/s(b_0) = 576.68/23.47 = 24.57,$$

  and the cut value

  $$t_{\alpha/2}(n-2) = t_{0.05/2}(30-2) = 2.04,$$

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

Hypothesis Test for Slope: Example

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

  $$T_0 = b_1/s(b_1) = -3.0068/0.4243 = -7.09,$$

  and the cut value

  $$t_{\alpha/2}(n-2) = t_{0.05/2}(30-2) = 2.04,$$

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

Two-Sided vs. One-sided Test

- Two-sided Test:

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

  One-sided Test:

$$H_0 : \beta_1 = 0, \; H_1 : \beta_1 > 0$$

- Two-sided Test: Reject H_0 when $|T_0| > t_{\alpha/2}(n-2)$ or $P(|T| > |T_0|) < \alpha$

  One-sided Test: Reject H_0 when $T_0 > t_\alpha(n-2)$ or $P(T > T_0) < \alpha$

Cont...

- Two-sided Test:

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

One-sided Test:

$$H_0 : \beta_1 = 0, \; H_1 : \beta_1 < 0$$

- Two-sided Test: Reject H_0 when $|T_0| > t_{\alpha/2}(n-2)$ or $P(|T| > |T_0|) < \alpha$
One-sided Test: Reject H_0 when $T_0 < -t_{\alpha}(n-2)$ or $P(T < T_0) < \alpha$

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 $1-\alpha$. 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 Intervals

- A generic format for CI is:

$$(\text{point estimate} \pm (\text{Multiplier } t^{\star} \times \text{sd of the point estimate}))\,.$$

  - **point estimate:** is the value that estimates the population parameter based on a random sample, e.g. b_1 for β_1, and b_0 for β_0.
  - **Multiplier $t^{\star}$:** depends on the confidence level and the distribution associated with the point estimate.
  - **Standard deviation (sd):** measures the accuracy of the point estimate.

Confidence Interval for Slope β_1

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

$$
\begin{aligned}
&(\text{point estimate} \quad \pm \quad \text{Multiplier } t^\star \times \text{sd of the point estimate})\\
&(b_1 \quad \pm \quad t_{\alpha/2}(n-2) \times s(b_1))\\
&(b_1 \quad \pm \quad t_{\alpha/2}(n-2) \times \frac{\sqrt{MSE}}{\sqrt{\sum_{i=1}^{n}}(x_i - \bar{x})}
\end{aligned}
$$

Interpretation: (Recall interpretation of b_0) One unit increase in the predictor will be associate with a change in the mean of the response. With $(1-\alpha)$ confidence, this change is somewhere in the interval $(b_1 - t_{\alpha/2}(n-2) \times s(b_1), b_1 + t_{\alpha/2}(n-2) \times s(b_1))$.

Q: What influence the width of the confidence interval for β_1?

Hypothesis Testing v.s. Confidence Interval

They are equivalent in some sense. Consider the hypothesis testing:

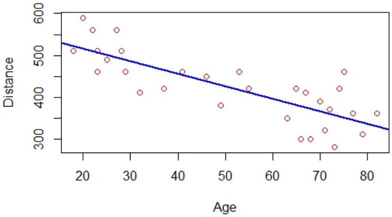
$$H_0 : \beta_1 = 0, \; H_1 : \beta_1 \neq 0$$

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

- If CI for β_1 contains 0, we conclude that there is no evidence of a linear relationship between the predictor and the response in the population
- If CI for β_1 does not contain 0, we conclude that there is evidence of a linear relationship between the predictor and the response in the population.

Confidence Interval for Slope β_1: Example

Example: $n = 30$ observations on driver age and the maximum distance (feet) at which individuals can read a highway sign:




```
Call:
lm(formula = Distance ~ Age, data = sign)

Residuals:
    Min      1Q  Median      3Q     Max
-78.231 -41.710   7.646  33.552 108.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
Age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,     Adjusted R-squared:  0.6292
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07
```

Confidence Interval for Slope β_1: Example

In our example, $n = 30$ and $\mathrm{df} = n - 2 = 28$. For 95% confidence, $t^\star = t_{(1-95\%)/2}(28) = 2.048$. A 95% confidence interval for β_1 , the true population slope is:

$$(-3.0068 \pm (2.048 \times 0.4243)) \approx (-3.88,\ -2.14).$$

Interpretation: With 95% confidence we can say the mean sign reading distance decreases somewhere between 2.14 and 3.88 feet with one-year increase in age.
Testing: We should reject the null hypothesis $H_0:\ \beta_1 = 0$ at 0.05 significance level, because the 95% CI doesn't contain 0.

Confidence Interval for Slope β_1: Example

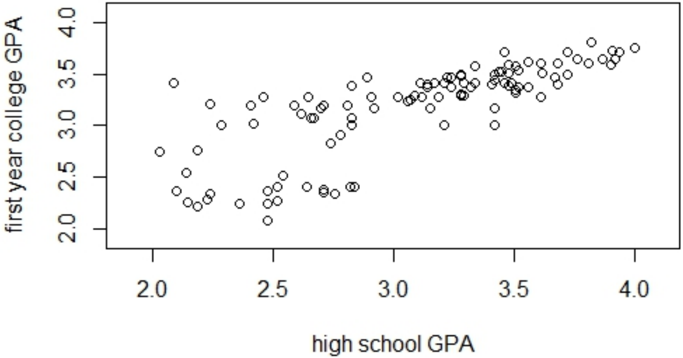
If we want to get 99% confidence interval,
$t^{\star} = t_{(1-99\%)/2}(28) = 2.763$. Then a 99% confidence interval estimate of β_1 is:

$$(-3.0068 \pm (2.763 \times 0.4243)) \approx (-4.18, -1.84).$$

Interpretation: With 99% confidence we can say the mean sign reading distance decreases somewhere between 1.84 and 4.18 feet with one-year increase in age.
NOTE: The same procedure can be used to calculate a confidence interval for the population intercept. Just use b_0 and its standard error rather than b_1. **Interpretation?**

Do college students keep their high school GPA?

high school GPA and first year college GPA of 105 students.



Do college students keep their high school GPA?

Some questions that admission officer may interest:

- What is the average first year GPA for applicants who have high school GPA 3.0?
- What is the most likely first year GPA for an applicant with high school GPA 3.0?

Suppose the officer fits a simple linear regression for this data set and get:

$$\text{Fitted first year GPA} = 1.10 + 0.675 \times \text{high school GPA}$$

Confidence Interval for $E(Y)$

- Question 1: What is the average first year GPA for applicants who have high school GPA 3.0?

$$\text{Fitted value? } \hat{y} = 1.10 + 0.675 \times 3 = 3.125$$

We can do better!

Confidence Interval for $E(Y)$

- **A 95% Confidence Interval** –a interval that we claim with 95% confidence the average first year GPA will in it!

$$
\begin{array}{ccc}
\text{(fitted value} & \pm & \text{Multiplier } t^{\star} \times \text{standard error of the fitted value)} \\
(\widehat{y} & \pm & t_{(1-0.95)/2}(n-2) \times s\{\widehat{y}\})
\end{array}
$$

- Suppose $s\{\widehat{y}\} = 0.0278$, $t_{0.025}(103) = 1.98$, then a 95% C.I is

$$3.125 \pm 1.98 * 0.0278 = (3.06, 3.18)$$

For applicants whose high school GPA are 3.0, with 95% confidence we can estimate the mean GPA in their first year college is between 3.06 and 3.18.

Confidence Interval for $E(Y)$

A $1-\alpha$ **confidence interval for** $E(Y)$ is an interval estimate for the mean value of y or, $E(Y)$ in the population level given an x.

$$(\widehat{y} \pm t_{(1-0.95)/2}(n-2) \times s\{\widehat{y}\})$$

where $\widehat{y} = b_0 + b_1 x$

- a confidence interval for $E(Y)$ estimates the location of the line at a specific x value.
- It counts the variation of different samples.
- C.I with higher confidence level will be wider. e.g. 99% C.I is wider than 95% C.I.
- $s\{\widehat{y}\} = \sqrt{MSE \cdot (\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2})}$

Prediction Interval for y

- Question 2: What is the most likely first year GPA for an applicant with high school GPA 3.0?

  Confidence interval?

  C.I are for the averages, we want an interval prediction of first year GPA for this applicant (an individual observation with $x_h = 3.0$).

Prediction Interval for y

A **prediction interval** is an interval estimate for a new observation y corresponding to a given level x_h.

- e.g. use a P.I. to predict first year GPA (y) for an individual with high school GPA 3.0 ($x_h = 3.0$)
- It has to count for the variation in the mean (as in confidence intervals), but also count for random error of observations(e.g. individual differences in GPA example).
- Wider than a confidence interval.
- Interpretation similar as a confidence interval.

Prediction Interval for y

- A $(1-\alpha)$ prediction interval for y give x_h is

$$\hat{y} \pm t_{\alpha/2}(n-2)\sqrt{s^2\{\hat{y}\}+MSE}$$

where $\hat{y} = b_0 + b_1 * x_h$.

- In the GPA example, $s\{\widehat{y}\} = 0.0278$, $t_{0.025}(103) = 1.98$, $MSE = 0.079$ then a 95% P.I. for y given $x_h = 3$ is

$$3.125 \pm 1.98 * \sqrt{0.0278^2 + 0.079} = (2.57, 3.68)$$

If the high school GPA of an applicant is 3.0, then we have 95% confidence to predict his/her first year college is between 2.57 and 3.68.

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

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	3.1213	0.0278	(3.0662, 3.1764)	(2.5604, 3.6822)

- "95% CI" of the mean of first year college GPA for applicants with high school GPA 3.0:(3.0662,3.1764)
- "95% PI" of first year college GPA for a new applicant with high school GPA 3.0:(2.5604,3.6822)
- "Fit":is calculated as $\hat{y} = 1.10 + 0.675 \times 3 = 3.12$.
- "SE Fit":
  is the standard deviation of $\hat{y}$ $(or, s(\hat{y}))$; it measures the accuracy of $\hat{y}$ as an estimate of $E(Y)$.

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

	object	formula
C.I.	mean $E(Y)$	$\hat{y} \pm t_{\alpha/2}(n-2) \times s\{\hat{y}\}$
P.I.	observation y	$\hat{y} \pm t_{\alpha/2}(n-2) \times \sqrt{MSE + s^2\{\hat{y}\}}$

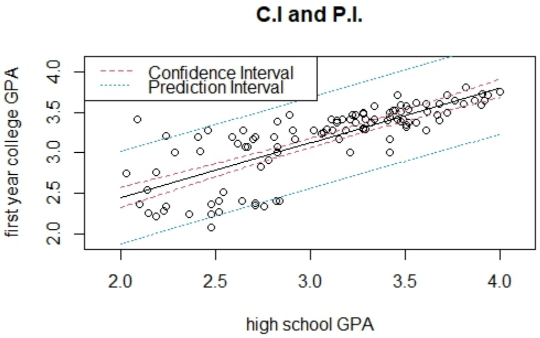
	interpretation
C.I.	with 95% confidence, we can estimate the mean of the response for a given x is in the C.I.
P.I.	with 95% confidence, we can predict the response value for a given x is in the P.I.

One last note

Attention!!

1. C.I. and P.I. can only be used when x is a value within the range of the x values in the data set.
    - e.g. GPA data set, we can only calculate C.I. and P.I. when high school GPA is between 2.03 and 4.00.
2. **Extrapolation**: when x is "out the scope of the model"
    - Don't know whether it still follows the same linear regression model.
    - Does not make sense.
3. But x does not have to be one of the actual x values in the data set.
    - e.g. GPA data set, we can calculate C.I. and P.I. for any number between 2.03 and 4.00.

Confidence Interval and Prediction Interval



Analysis of Variance (ANOVA)

```
Residuals:
    Min      1Q  Median      3Q     Max
-78.231 -41.710   7.646  33.552 108.831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
Age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,     Adjusted R-squared:  0.6292
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07

> anova(out2)
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Two uses of ANOVA table

1. A decomposition of variance in y
2. A significance test of whether x and y are really linearly related in the population.

1. A decomposition of variance in y

```
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)    
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Main Quantities:

1. Sum of Squares due to Regression (SSR)=124333.
   The variation of drivers' abilities to read highway sign due to their age differences.
2. Sums of Squared Errors (SSE)=69334.
   The variation of drivers' abilities to read highway sign which can not explained by their ages differences.(other individual difference)
3. Sums of Squares for Total (SST)=193667.
   How much the observed drivers' reading distances vary if you don't take into count their age differences.

1. A decomposition of variance in y

```
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

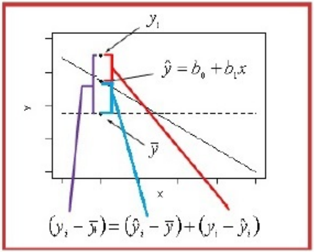
Main Relations:

1. in SS $193667 = 124333 + 69334$ $\Leftrightarrow$SST=SSR+SSE

$$\begin{aligned} & \text{Overall Variation in } y \text{ variable}(SST) \\ = & \text{ Variation “due to” change of X}(SSR) \\ & +\text{Variance “due to” random error}(SSE) \end{aligned}$$

2. in DF 29=28+1 $\Leftrightarrow$df(SST)=df(SSR)+df(SSE)

1. A decomposition of variance in y



$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\boxed{\text{SST} \quad = \quad \text{SSR} \quad + \quad \text{SSE}}$$

$$\text{df(SST)} = \text{df (SSR)} + \text{df(SSE)}$$

simple linear regression: $(\,n-1\,) = 1 + (n-2)$

- SST: quantifies how much the observed responses vary if you don't take into account their predictor values.
- SSR: it quantifies how far the estimated regression line is from the no relationship line.
- SSE: it quantifies how much the data points vary around the estimated regression line.

2. A significance test

```
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Mean Square for the Regression (MSR):

$$MSR = SSR/df(SSR) = SSR/(1), \text{ e.g. } 12344 = 12344/1$$

- Mean Squared Error(MSE):(estimates of σ^2)

$$MSE = SSE/df(SSE) = SSE/(n-2), \text{ e.g. } 2476 = 69334/28$$

- F statistics: used to test the significance in the linear relation

$$F = MSR/MSE, \text{ e.g. } 50.21 = 12344/2476$$

2. A significance test

F statistic to test whether the y-variable and x-variable are related:

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0$$

And in simple linear regression $p = 2$

$$F = \frac{MSR}{MSE} \sim F(p-1, n-p), \text{ under } H_0 : \beta_1 = 0$$

1. **Critical value approach:** Reject H_0 if the calculated statistic F_0,
$$F_0 > F_\alpha(1, n-1),$$
where $F_\alpha(1, n-1)$ is the $1-\alpha$ percentile of $F(1, n-1)$ distribution, α usually takes as 0.05.

2. **p-value approach:** Reject H_0 if $p - value \leq 0.05$

2. A significance test

```
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value of F-test is $0.000 < 0.05$, so we have enough evidence to reject the null hypothesis. This suggest that we have enough evidence to say that there is indeed linear relation between a driver's age and his/her maximum distance to read a highway sign.

t-test v.s. F-test

Do we have two tests of significance??

```
Residuals:
    Min      1Q  Median      3Q     Max 
-78.231 -41.710   7.646  33.552 108.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
Age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07

> anova(out2)
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)    
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. Both test for linear relationship or $H_0 : \beta_1 = 0$ v.s.$H_0 : \beta_1 \neq 0$
2. $[t(n-1)]^2 = F(1, n-1)$, so $(t - value)^2 = (F - value)$
   e.g. $t_0^2 = (-7.09)^2 = 50.2 = F_0$
3. They have exactly the same p-values.

t-test v.s. F-test

Difference:

- F can only be used for two-sided tests, t can also be used for one-sided tests.
- When to use F?
  For **multiple regression**, while t-test is used to test the significance of each β coefficient, F-test is used to test all coefficients simultaneously, i.e.:

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

Analysis of Variance (ANOVA): Table

Source	DF	SS	$MS = SS/DF$	F
Regression	1	SST-SSE	MSR	MSR/MSE
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		