We did regression \( y_i = \alpha + \beta x_i + \cdots + \epsilon_i \).  
We did inference on \( \mu, \pi, \bar{y} - \mu, \bar{x} - \bar{y}, \bar{y}, \cdots \).  
Now we do inference on \( \beta, \) (and \( \alpha \)), \( y \), \cdots.

Review:\( \hat{\gamma}_i = y_i - \hat{\gamma} \hat{x}_i \)  
\( L = \hat{\alpha} + \hat{\beta} x + \cdots \)

For a sample we write \( y_i = \hat{\alpha} + \hat{\beta} x_i + \epsilon_i \) arbitrary params to be estimated by OLS, i.e., \( \frac{\partial}{\partial \alpha} \text{SSE}, \text{etc.} \)  
where \( \hat{\alpha}, \hat{\beta} \) are the OLS estimates of \( \alpha, \beta \), i.e.,  
\( \hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \)  
\( \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \),  
where \( S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \)  
\( S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \)  
Recall that \( \text{Sample var.} = s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{S_{xx}}{n-1} \)

For a population, there exists an OLS fit as well!  
Just because we can fit a line to the whole pop., it does not follow that there are no errors when predicting \( y \) from \( x \).

So, now, in regression we need notation that can distinguish between pop. and sample quantities; like \( \bar{x} \), \( \bar{y} \), \( \mu_x \), but for regression, I will use the following notation for the predictions:  
\( \hat{y}(x) = \hat{\alpha} + \hat{\beta} x \) (for sample)  
\( y(x) = \alpha + \beta x \) (for population)  
But, in CH11, you have to keep in mind that this \( \alpha, \beta \) are NOT free params that we can do things like \( \frac{\partial}{\partial \alpha} \text{SSE}, \text{etc.} \); they are fixed quantities obtained by "fitting" a line to the whole population.
Then there is the analysis of variance:

\[ \text{SST} = \sum \left( y_i - \bar{y} \right)^2 = \text{SS explained} + \text{SS unexplained} \]

\[ \hat{b} = \frac{\text{SS explained}}{\text{SST}} \]

\[ R^2 = \frac{\text{SS explained}}{\text{SST}} \]

- percent of variance explained by x
- goodness of fit

\[ \text{SSE} = \sum (y_i - \hat{y}_i)^2 \]

\[ \text{SST} = \sum (y_i - \bar{y})^2 \]

\[ \text{df} = n - 1 \]

\[ \text{df} = \text{df for } \hat{b} \]

\[ \text{df} = \text{df for S.E.} \]

\[ \text{df} = n - (k + 1) \]

\[ s_e = \sqrt{\frac{\text{SSE}}{n - (k + 1)}} \]

- standard deviation of errors
- typical error or spread about fit

Now, to do inference we need a probability model (for regression):

Assume y's are normally distributed at each x, with means \( \mu = \gamma(x) \), \( \sigma = \sigma_e \)

\[ \gamma(x) = \alpha + \beta x + \varepsilon \]

\( \alpha \), \( \beta \) estimates, \( \hat{\alpha} \), \( \hat{\beta} \)

\[ \varepsilon = y - \gamma(x) \sim \text{N}(0, \sigma_e) \]

\[ y \sim \text{N}(\gamma(x), \sigma_e) \]

This allows us to say things like:

1) \( \hat{\gamma}(x) = \hat{\alpha} + \hat{\beta} x \) = estimates mean of y, given x

2) In about 95% of the cases, we expect to have y-values within \( \gamma(x) \pm 1.96 \sigma_e \), for a given x

3) Other probs, e.g.

\[ \text{prob}(a < y < b | x) = \]

\[ \frac{1}{2} \text{ prob} \left( \frac{a - \gamma(x)}{\sigma_e} < \frac{y - \gamma(x)}{\sigma_e} < \frac{b - \gamma(x)}{\sigma_e} \right) = \text{Table I} \]

\[ \frac{a - \mu}{\sigma} \sim \text{N}(0, 1) \]

\[ \frac{b - \mu}{\sigma} \sim \text{N}(0, 1) \]

\[ \text{prob} (a < x < b) = \text{prob} (a < \hat{y} < b | x) \]

In short: For a fixed x, everything we have done (CI, p-value...) now applies to y.

Like 95% of cases are within \( x \pm 1.96 \sigma \) (Ch. 1).
Let's build a CI (and hyp. test) for ONE $\beta$: $Y_i = \alpha + \beta \times_i + \varepsilon_i$

**Theorem:** If $\varepsilon \sim N(0, \sigma^2_\varepsilon)$, then $\hat{\beta}$ is normal with means:

- $E[\hat{\beta}] = \beta$ (pop. slope)
- $\sqrt{\text{Var}[\hat{\beta}]} = \frac{\sigma_\varepsilon}{\sqrt{S_{xx}}} = \frac{\sigma_\varepsilon}{\sqrt{n-1} S_x}$

**Ch. 7**

- If $x \sim N(x, \sigma^2_x)$, then $\bar{x}$ is normal with means $E[\bar{x}] = \mu_x$ and $\sqrt{\text{Var}[\bar{x}]} = \frac{\sigma_x}{\sqrt{n}}$

**Recall** $\sigma_\varepsilon$ is const., and $S_x$ does not vary as $\frac{1}{n-1}$ because of 2 in the numerator if $S_x$.

**Q1:** What is the quantity that has a std. normal dist.?

- A) $\frac{\hat{\beta} - \beta}{\sigma_\beta / \sqrt{n}}$
- B) $\frac{\hat{\beta} - \beta}{\sigma_\beta / \sqrt{n}}$
- C) $\frac{\hat{\beta} - \mu_y}{\sigma_y / \sqrt{n}}$
- D) $\frac{\hat{\beta} - \mu_y}{\sigma_y / \sqrt{n}}$

If $w \sim N(\mu_w, \sigma_w^2)$, then $z = \frac{w - \mu_w}{\sigma_w}$ follows $N(0,1)$.

$z = \frac{\hat{\beta} - \beta}{\sigma_\varepsilon / \sqrt{S_{xx}}} \sim N(0,1)$

$\therefore t = \frac{\hat{\beta} - \beta}{\sigma_\varepsilon / \sqrt{S_{xx}}} \sim t$-dist., $df = n - 2$

Then, self-evident fact gives:

**C.I. for $\beta$**:

$\hat{\beta} \pm \frac{t \cdot \sigma_\varepsilon}{\sqrt{S_{xx}}} \quad df = n - 2$ (Table VI)

**H_0**: $\beta = \beta_0$

$H_1$: $\beta \neq \beta_0$

$t_{obs} = \frac{\hat{\beta}_{obs} - \beta_0}{\sigma_\varepsilon / \sqrt{S_{xx}}} \quad df = n - 2$ (k+1)

**P-value** = $(1,2) \cdot p\text{-value}(\beta \neq \beta_0|t_{obs}) = p\text{-value}(t \neq t_{obs}) = \text{Table VI}$

$\leq 1$ or 2-sided.
Problem 11.17 [Revised; remove the word "positive", i.e. do 2-sided]

\(n=13\) \(x = \) nickel content, \(y = \) percentage austenite.

Data:
\[
\begin{align*}
\sum (x_i - \bar{x})^2 &= 1.183 &= S_{xx} \\
\sum (y_i - \bar{y})^2 &= 0.0508 &= S_{yy} \\
\sum (x_i - \bar{x})(y_i - \bar{y}) &= 0.2073 &= S_{xy}
\end{align*}
\]

Question: Is there a statistically significant \((\alpha = 0.05)\) relationship between \(x\) and \(y\)?

1) CI for \(\beta\):
\[
\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{0.2073}{1.183} = 0.1752
\]
\[
S_{e} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.14}{13-2}} = 0.0357
\]
\[
S_{e} = 0.0357
\]

\[
\text{95\% CI for } \beta: \quad 0.1752 \pm 2.201 \left( \frac{0.0357}{\sqrt{1.183}} \right) = (0.10, 0.24)
\]

We are 95\% confident that the pop. \(\beta\) is in here.
Also, zero is not included \(\Rightarrow\) Relationship is statistically significant.

2) \(H_0: \beta = 0\)
\(H_1: \beta \neq 0\)

\[
t_{\text{obs}} = \frac{0.1752 - 0}{0.0328} = 5.31
\]

\[
p\text{-value} = 2 \cdot p(t > t_{\text{obs}}) = 2 \cdot p(t > 5.31) = 2 \cdot 0.001 = 0.002
\]

\(p\)-value < \(\alpha\)

\(\Rightarrow\) Evidence that \(\beta \neq 0\). (Same conclusion as above.)

In summary, we have 2 ways of testing an association between \(xy\).

(A third way, next FYI)
Note that the test of $\beta = 0$ is equivalent to testing if there is a linear relationship between $x$ and $y$. But if a linear relationship is all that you are testing, then we can test the population correlation coeff

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

the test statistic for this test is a bit weird:

$$t = \frac{r - 0}{\sqrt{1 - r^2} \sqrt{\frac{n-2}{n-2}}}$$

has a $t$ distrib, with $df = n-2$.

Recall $r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$

This way, you take your data $(x_i, y_i)$, compute the sample correlation coeff $(r)$, then $t$ obs, and then $p$-value, all without any fitting.

3) For the above example:

$$H_0: \rho = 0 \quad r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \cdots = 0.8456$$

$$t_{obs} = \frac{r - 0}{\sqrt{1 - r^2} \sqrt{\frac{n-2}{n-2}}} = \cdots = 5.3 < \text{some value as } t_{obs} \text{ we got above when testing } \beta.$$ 

$p$-value = $2 \cdot \text{prob} (t > t_{obs}) = \text{same as above}$. Some conclusion.

In summary: We have 3 ways of testing if there is a useful relation between $x$ & $y$:

1) C.I. for $\beta$ 2) Testing $H_0: \beta = 0$ 3) $H_0: \rho = 0$
The very beginning of section 3.3 in lab4 shows how to make/simulate data on x and y that are linearly associated. The x data consists of 100 cases from a uniform distribution, and the TRUE/population relationship between x and y is given by \( y = 10 + 2x \).

a) What is the value of sigma_epsilon in that simulation?

b) Using the same settings used in section 3.3, write code to build the (empirical) sampling distribution of beta_hat based on 5000 trials. This code should produce a histogram.

c) According to the lecture, the mean of the histogram is supposed to be equal (or close) to what quantity? Is it?

d) According to the lecture, the standard deviation of that histogram is supposed to be equal (or close) to what quantity? Is it?

e) According the lecture, the distribution of the beta_hat is supposed to be normal with certain parameters. Use qqnorm() and abline() to confirm that.

In a problem dealing with flow rate (y) and pressure drop (x) across filters, it is known that \( y = 0.12 + 0.095x \). I.e. This is the “fit” to the population. Suppose it is also known that \( \sigma_e = 0.025 \). Now, IF we were to make repeated observations of y when x=10, what’s the prob. of a flow rate exceeding 0.835?

Hint: \( y - \left( \text{true mean of } y \text{ at some } x \right) \sim N(0, \sigma_e) \).