We did regression \( y_i = \alpha + \beta x_i + \cdots + \epsilon_i \) \( \text{Ch. 3.} \)
We did inference on \( \mu, \tau, \mu_1 - \mu_2, \bar{y}_1 - \bar{y}_2, \bar{y}_i, \cdots \) \( \mu = \mu_1 = \cdots \) \( \text{Ch. 7, 8, 9} \)
Now we do inference in regression (on \( \beta, \alpha, \gamma, \cdots \)) \( \text{Ch. 11.} \)

Review:

\[ \hat{y}_i = \hat{\beta} \hat{x}_i + \cdots \]
\[ \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \]
\[ \hat{\beta} = \frac{S_{xy}}{S_{xx}} \]
\[ \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \]

For a sample we write
\[ y_i = \alpha + \beta x_i + \epsilon_i \]
\[ \text{in book} \]
where \( \hat{\beta}, \hat{\alpha} \) are the OLS estimates of \( \beta, \alpha \), ie.

\[ \frac{\hat{\beta}}{\hat{\alpha}} = \text{SS}_E \text{, etc.} \]

Recall that
\[ \text{Sample Var. } = s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = \frac{S_{xx}}{n-1} \]

There is also the Analysis of Variance:

\[ \text{SST} = \sum (y_i - \bar{y})^2 = \text{SS explained} + \text{SS unexplained} \]
\[ \frac{\hat{\beta} S_{xy}}{S_{xx}} \]
\[ \text{df } = n-1 \]
\[ k + n - (k+1) \]

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

percent of variance explained by \( x \) (excluding \( \alpha \) for predictive use)

\( \text{Goodness of fit} \) \( y = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \)
\[ \text{MSE} = \frac{\text{SSE}}{n - (k + 1)} \]

\[ \text{Typical error} \text{ or spread about fit.} \]
For population

Now, let's consider the population. For the moment, suppose we have it. Just because we have the population, it does not follow that there is no scatter between x, y. i.e. even for the population, there is a scatter between x and y, and so there is an OLS fit for the population! i.e. even for the population, there is an OLS fit! Call it the true fit. What symbols should we use to denote this true fit?

<table>
<thead>
<tr>
<th>Sample mean: $\bar{x}$</th>
<th>Sample OLS fit: $\hat{a}$, $\hat{b}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True/pop/distr. mean: $\mu_x$</td>
<td>True/pop/distr. OLS fit: $\mu$</td>
</tr>
</tbody>
</table>

It's tempting to use $\alpha, \beta$ (w/o hat). But $\alpha, \beta$ are supposed to be free parameters to minimize the SSE.

So, technically, we should introduce new symbols for the true OLS fit. However, to keep things simple, we will go ahead and use $\alpha, \beta$ to denote the true OLS fit.

I.e. up to now, $\alpha, \beta$ have been free parameters to do $\hat{y} = \hat{a}x, \hat{y} = \hat{b}$. But henceforth, $\alpha, \beta$ denote the OLS fits for the population.

In short: $\hat{y}(x) = \hat{a} + \hat{b}x$ (for sample) and $\hat{y}(x) = \alpha + \beta x$ (for population)

will be used for the respective predicted values.

I.e. $\hat{y}(x) =$ prediction (in sample)

$\hat{y}(x) =$ prediction (in pop.)
Now, to do inference we need a probability model (for regression):

Assume y's are normally distr. at each x, with params $\mu = \gamma(x)$, $\sigma = \sigma_\epsilon$

\[ e.g. \quad \mu = \gamma(x) = \alpha + \beta x + \ldots \quad \sigma = \sigma_\epsilon = \text{fixed} \]

estimate $\alpha, \beta$ with $\hat{\alpha}, \hat{\beta}$ estimate with $\hat{\epsilon}$

Note: $\gamma \sim \mathcal{N}(\gamma(x), \sigma_\epsilon)$

$\epsilon = y - \gamma(x) \sim \mathcal{N}(0, \sigma_\epsilon)$

This allows us to say things like:

1) $\hat{\gamma}(x) = \hat{\alpha} + \hat{\beta} x$ estimates true mean of $y$, given $x$ i.e. $\gamma(x)$.

2) At a given $x$, we expect about 95% of the $y$'s to be within $\gamma(x) \pm 1.96 \sigma_\epsilon$.

3) Other probs, e.g. $\text{prob}(a < y < b | x) = \text{True prediction } = \text{True mean } x.$

\[ \text{prob}(\frac{a - \gamma(x)}{\sigma_\epsilon} < \frac{y - \gamma(x)}{\sigma_\epsilon} < \frac{b - \gamma(x)}{\sigma_\epsilon}) = \text{Table I} \]

$\quad \mathcal{Z} \sim \mathcal{N}(0,1)$

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

Like 95% of $x$'s are within $\mu \pm 1.96 \sigma$ (Ch. 1)

because $\not x \not x$
n = 10
n.trial = 64

x = c(1:n)
y.true = 10 + 2*x
sigma_eps = 15

Note that the x-values are the same across trials. (In the kind of regression we are doing, x has no uncertainty; only y does.)
Let's build a CI (and hyp. test) for ONE $\beta$ : 

$$\gamma_i = \alpha + \beta x_i + \epsilon_i$$

**Theorem:** If $\epsilon \sim N(0, \sigma_\epsilon^2)$, Then $\hat{\beta}$ is normal with

$$E[\hat{\beta}] = \mu_{\hat{\beta}} = \beta \quad \text{pop. slope}$$

$$\sqrt{\text{Var}[\hat{\beta}]} = \sigma_{\hat{\beta}} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}} \quad \text{(not obvious)}$$

$$\text{Ch. 7:}$$

- If $x \sim N(\mu_x, \sigma_x^2)$, Then $\bar{x}$ is normal with
  
  $$E[\bar{x}] = \mu_{\bar{x}} = \mu_x$$

  $$\sqrt{\text{Var}[\bar{x}]} = \sigma_{\bar{x}} = \sigma_x/\sqrt{n}$$

$$\Rightarrow S_{xx} = \sum \frac{1}{n} (x_i - \bar{x})^2 = (n-1) S_x^2$$

- **Defn. of sample var.**

Since $\hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}}^2)$, Then

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

$$\Rightarrow t = \frac{\hat{\beta} - \beta}{\sigma_\epsilon/\sqrt{S_{xx}}} \sim t - \text{dist.}$$

**df = n-2, k+1**

**Ch. 7:**

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

$$t = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim t - \text{dist.}$$

**df = n-1**

Then, self-evident fact gives:

C.I. for $\beta$:

$$\hat{\beta} \pm z^{*} \frac{\sigma_\epsilon}{\sqrt{S_{xx}}} \quad \text{df = n-2 (Table VII)}$$

$H_0 : \beta = \beta_0$

$$t_{obs} = \frac{\hat{\beta} - \beta_0}{\sigma_\epsilon/\sqrt{S_{xx}}} \quad \text{df = n-2, k+1}$$

$H_1 : \beta \neq \beta_0$

$$p-value = (1, 2). FV(\hat{\beta} \bigcap \beta_{obs}) = FV(t \bigcap t_{obs}) \quad \text{Table VII}$$

$\leq \alpha$ or 2-sided.
Problem 11.17 [Revised]

n=13  x = nickel content, y = percentage austenite.

Data: $E(x_i - \bar{x})^2 = 1.183 = S_{xx}$
$E(y_i - \bar{y})^2 = 0.0508 = S_{yy} = \text{SST}$
$E(xy - \bar{x}\bar{y}) = 0.2073 = S_{xy}$

Question: Is there a statistically significant ($\alpha=0.05$) relationship between x and y? Hint: $SS_{exp} = \hat{\beta} S_{xy}$

CI: $\beta: \hat{\beta} \pm t^* \frac{S_e}{\sqrt{S_{xx}}}$

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1.2073}{1.183} = 1.0752, \quad \text{SSE} = \text{SST} - \text{SS}_{exp} = 0.0508 - (1.0752)(0.2073) = 0.14$

$S_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{0.14}{12}} = 0.0357$

$95\% \ CI \ for \ \hat{\beta}: \hat{\beta} \pm t_{0.025} \left( \frac{S_e}{\sqrt{S_{xx}}} \right) = 1.0752 \pm \frac{2.201(0.0357)}{\sqrt{1.183}} = (0.10, 0.24)$

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_{obs} = \frac{1.0752 - 0}{0.0328} = 3.31$

$p-value = 2 \cdot P(t > t_{obs}) = 2 \cdot P(t > 3.31) = 2 \cdot 0.001 < 0.05$

$p-value < \alpha$

Evidence that $\beta \neq 0$. (same conclusion as above).

In summary: we have 2 ways of testing whether there is a relationship between 2 continuous variables.
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:

$$t = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n-2}}}$$  has a t distv. 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 correl. coeff ($r$), then $t obs$, and then p-value, all without any fitting.

3) For the above example:

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$r = \frac{\sum{xy}}{\sqrt{\sum{x^2} \sum{y^2}}} = \ldots = 0.8456$

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

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

i. Same conclusion.
hw_lect23_1: By R
a) Revise the simulation shown in the lecture with the aim of constructing the empirical sampling distribution of beta_hat, based on 5000 trials.

b) According to the lecture, the mean of that histogram is supposed to be approximately equal to the true slope. Is it? Show code.

c) According to the lecture, the standard deviation of that histogram is supposed to be approximately equal to sigma_eps/sqrt(Sxx). Is it? Show code.

d) According to the lecture, the distribution of the beta_hat is supposed to be normal with parameters found in parts a and b. Use qqnorm() and abline() to confirm that it is normal.

hw_lect23_2
In a problem dealing with flow rate (y) and pressure-drop (x) across filters, it is known that $y = -0.12 + 0.095 x$. Note: this is the true "fit" to the population. Suppose it is also known that sigma_epsilon = 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?