Lecture 23 (Ch. 11)

We did regression \( y_i = \alpha + \beta x_i + \ldots + \epsilon_i \) \( \text{Ch. 3} \). We did inference on \( \mu, \mu_1, \mu_2, \mu_1 - \mu_2, \mu_i, \ldots \) \( \mu_i = \mu, \mu_i = \mu_i - \mu \) \( \text{Ch. 7, 8, 9} \).

Now, we do inference in regression (on \( \beta, \alpha, y(x), \ldots \)) \( \text{Ch. 11} \).

Review: \( y_i = \alpha + \beta x_i + \epsilon_i \). \( \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \).

For a sample we write \( y_i = \hat{\alpha} + \hat{\beta} x_i + \epsilon_i \).

\( \frac{\partial}{\partial \alpha} \text{SSE} \), etc.

\( \hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \).

where \( \hat{\alpha}, \hat{\beta} \) are the OLS estimates of \( \alpha, \beta \), i.e.

\( S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \)
\( S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \).

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

There is also the Analysis of Variance:
\( \text{SST} = \sum (y_i - \bar{y})^2 = \text{SS explained} + \frac{\text{SS explained}}{\text{SSE}} \).

\( \text{df} = n-1 \)
\( k = \# \text{ of } \beta \)’s
\( \text{df} = n - (k+1) \leftarrow (\text{EYI}) \)
\( R^2 = \frac{\text{SS explained}}{\text{SST}} \).

percent of variability in \( y \) explained by \( x \)...
\( \text{Goodness of fit} \)
\( \text{(excluding} \ \alpha \text{)} \)
\( \text{Std. dev. of errors} \ 
\frac{\text{SSE}}{\sqrt{n-(k+1)}} \text{NRMSE} \)
\( \text{Typical error or spread about fit} \).
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 pop.

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</th>
<th>( \bar{x} )</th>
<th>Sample OLS fit</th>
<th>( \hat{\alpha}, \hat{\beta} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>True/pop./distr. mean</td>
<td>( \mu_x )</td>
<td>True/pop./distr. OLS fit</td>
<td>( ? )</td>
</tr>
</tbody>
</table>

It's tempting to use \( \alpha, \beta \) (w/o hat). But \( \alpha, \beta \) are supposed to be free parameters that are tuned 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 \( \frac{\partial}{\partial \alpha}, \frac{\partial}{\partial \beta} \), etc. But henceforth, \( \alpha, \beta \) denote the OLS fits for the population.

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

will be used for the respective predicted values.

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

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

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

**e.g.** $\mu = y(x) = \alpha + \beta x + \varepsilon$  

$\sigma = \sigma_x$ = **fixed**  

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

**estimate with $\varepsilon_x$**  

**Note:** At a given $x$, $y \sim N(y(x), \sigma_x)$  

$\varepsilon = y - y(x) \sim N(0, \sigma_x)$  

This allows us to say things like:  

1) **At a given $x$, $\hat{y}(x)$ estimates the true mean of $y$, $y(x)$.**  

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

3) **At a given $x$, we can find other probs for $y$.**  

**e.g.** $\text{prob}(a < y < b | x) =$  

**True prediction = True mean $x$.**  

$\text{prob} \left( \frac{a-y(x)}{\sigma_x} < z < \frac{b-y(x)}{\sigma_x} \right) = Table I$  

$z \sim N(0, 1)$  

Note that these are just prob calculations, not p-values or CIs.
n = 10
n.trial = 64

x = c(1:n)
y_true = 10 + 2*x
sigma_eps = 15

par(mfrow=c(8,8),mar=c(0,0,0,0))
set.seed(123)
for(trial in 1:n.trial){
y_obs = y_true + rnorm(n,0,sigma_eps)
lm.1 = lm(y_obs ~ x)
plot(x, y_obs)
abline(10,2, col=2)
abline(lm.1, col=4)
}

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 $\beta$ : $y_i = \alpha + \beta x_i + \epsilon_i$

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

- $E[\hat{\beta}] = \mu_{\hat{\beta}} = \beta$ (pop. slope)
- $\sqrt{\text{V}[\hat{\beta}]} = \sigma_{\hat{\beta}} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}}$ (not obvious)

\[ \sigma_{\hat{\beta}} = \frac{\sigma_\epsilon}{\sqrt{S_{xx}}} \]
\[ S_{xx} = \sum (x_i - \bar{x})^2 = (n-1) S_x^2 \]
\[ \bar{\epsilon} = \frac{1}{n} \sum \epsilon_i \]

Since $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2_{\hat{\beta}})$, Then

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

\[ t = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} \sim t_{-dist}\]
\[ \text{df} = n-2 \]

Then, from self-evident fact

\[ \text{pv}(t < t^*) = \text{Conf. level} \]
\[ \text{df} = n-2 \text{ (Table VI)} \]

C.I. for $\beta$ : $\hat{\beta} \pm t^* \frac{\sigma_\epsilon}{\sqrt{S_{xx}}}$

$H_0 : \beta = \beta_0$

$t_{obs} = \frac{\hat{\beta}_{obs} - \beta_0}{\sigma_{\hat{\beta}}}$

$H_1 : \beta \neq \beta_0$

$p$-value = (1,2): $\text{pv}(\hat{\beta} | \beta_0) = \begin{cases} \frac{1}{2} & \text{1 or 2-sided} \\ \frac{2}{12} & \text{1 or 2-sided} \\ \text{(1,2) pv(} \pm | \beta_0 \text{)} = \begin{cases} \frac{1}{2} & \text{1 or 2-sided} \\ \text{Table VI, df} = n-1 \end{cases}$
Problem 11.17 [Revised]

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

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

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

1) C.I. \( \beta \):
\[
\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{0.2073}{1.183} = 0.1752 \quad \Rightarrow \quad SE = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.0508 - (0.1752)^2}{13-2}} = 0.0357
\]
\[\therefore \text{95% CI for } \beta: \quad 0.1752 \pm 2.201 \left( \frac{0.0357}{\sqrt{1.183}} \right) = (0.10, 0.24) \quad \text{df=13-2} \]

2) We are 95% confident that the pop. \(\beta\) is in here.
3) There is a 95% prob. that a random CI will cover \(\beta\).
4) Corollary: Relationship is statistically significant \((\text{zero not in CI})\).

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

\[ \text{t obs} = \frac{1.183 - 0}{0.0328} = 5.31 \quad \text{p-value} = 2 \cdot \text{pr}(t > t_{obs}) = 2 \cdot \text{pr}(t > 5.31) \]

\[= 2 \cdot \text{pr}(t > 5.31) < 0.001 \quad \text{df=13-2} \]

\[ \therefore \text{Evidence that } \beta \neq 0. \text{ (same conclusion as above) Table VI}. \]

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 $y$. 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 $t$ distribution 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 $t$-test, and then p-value, all without any fitting.

3) For the above example:

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

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

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

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

Same conclusion.
1. By R

a) Revise the simulation shown in the lecture with the aim of constructing the empirical sampling distribution of \( \beta_{\text{hat}} \), 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 \( \sigma_{\epsilon}/\sqrt{S_{xx}} \). Is it? Show code.

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

2. In a problem dealing with flow rate (y) and pressure-drop (x) across filters, it is known that

\[ y = -0.12 + 0.095x \]

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?