STA 302 / 1001

Note Title 9/29/2011

Model: $Y_i = \beta_0 + \beta_1 X_i + e_i$, $i = 1, \ldots, n$

Gauss-Markov conditions

$E(e_i) = 0$

$E(e_i e_j) = 0$, $i \neq j$

$Var(e_i) = \sigma^2$ for all i

+ Distributional assumption: e_i's have a Normal distn

So e_i's iid $N(0, \sigma^2)$

So Distribution of $Y_i \mid X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

$$\left(\hat{\beta}_1 \mid X_1 = x_1, \ldots, X_n = x_n\right) \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right)$$

So the minimum variance unbiased linear estimator of β_1 has a normal distribution

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) Y_i}{SXX}$$

$$SXX = \sum (x_i - \bar{x})^2$$

Note: $\left(\hat{\beta}_1 \mid X_1 = x_1, \ldots, X_n = x_n\right)$ is normally distributed

Estimate $\text{Var}(\hat{\beta}_1 | X)$ by $\frac{s^2}{S_{XX}}$ where $s^2 = \frac{\sum \hat{e}_i^2}{n-2}$

standard error - estimate of s.d. of a parameter

$$\text{s.e.}(\hat{\beta}_1 | X) = s / \sqrt{S_{XX}}$$

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{XX}}} \sim t_{n-2}$$

Confidence interval for $\hat{\beta}_1$

$100(1-\alpha)\%$ CI for β_1

$$b_1 \pm t_{\alpha/2,\, n-2} \frac{s}{\sqrt{S_{xx}}}, \quad s = \sqrt{MSE}$$

Similarly, $100(1-\alpha)\%$ CI for β_0

$$b_0 \pm t_{\alpha/2,\, n-2}\, s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

<u>Pre-MP data CFC data:</u>

90% CI for β_1 $(9.63, 9.80)$

Wider ↓

95% CI for β_1 $(9.61, 9.81)$

A 99% CI for β_1 (9.58, 9.85)

Common test in Simple Linear Regression

Want to test: $H_0: \beta_1 = 0$

vs $H_a: \beta_1 \neq 0$

Practically, is there a relationship between X & Y?

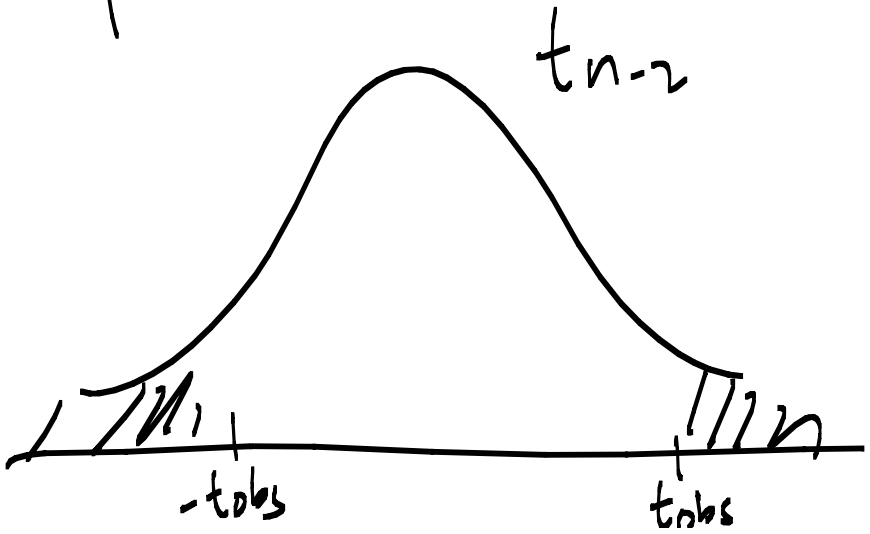
$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t_{n-2}$$

Test statistic for test $H_0: \beta_1 = 0$

$$t_{obs} = \frac{b_1}{s/\sqrt{SXX}}$$

If H_0 is true, t_{obs} is an observation from a t_{n-2} distribution.

Measure strength of evidence against H_0 with p-value



shaded area = p-value

Pre-MP CFC data $b_1 = 9.71152$

$S.e.(\hat{\beta}_1) = .05184$

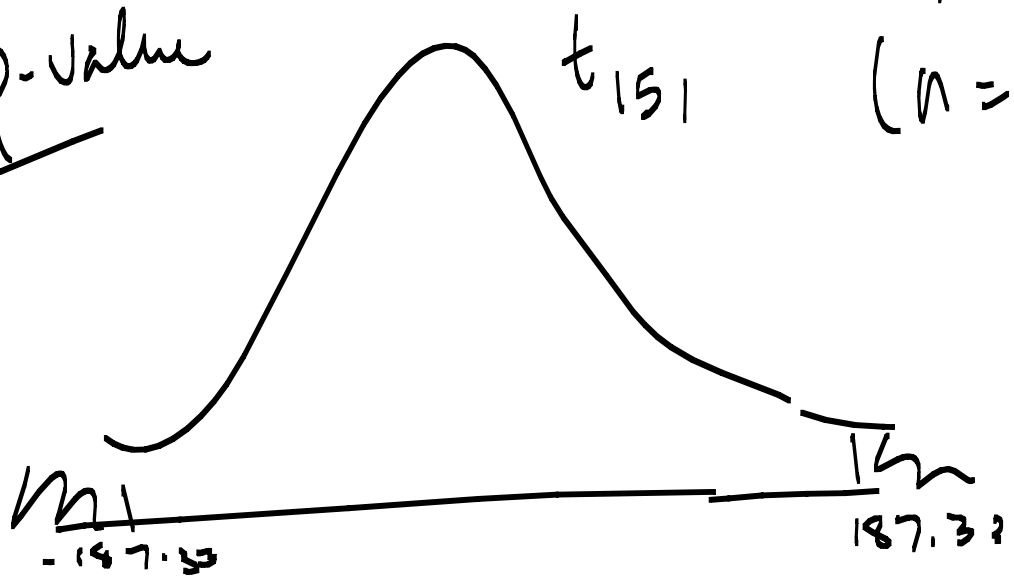
Test statistic for $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$H_a: \beta_1 > 0$

$$t_{obs} = \frac{9.71152}{.05184} = 187.33$$

$= 187.33$

p-value t_{151} $(n = 153)$



extremely small p-value

$p < .0001$

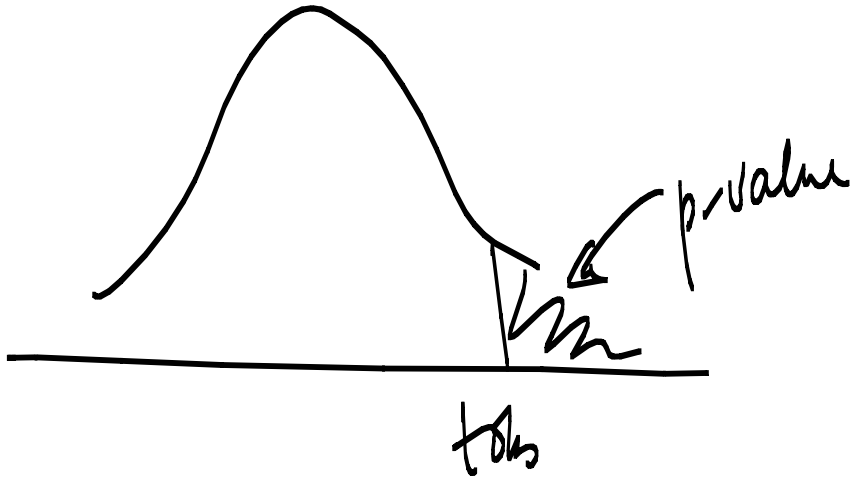
Very strong evidence that slope is different from 0

95% CI for slope was $(9.61, 9.81)$ $\Rightarrow$ know that 2-sided test for $\beta_1 = 0$ will have a p-value $< .05$

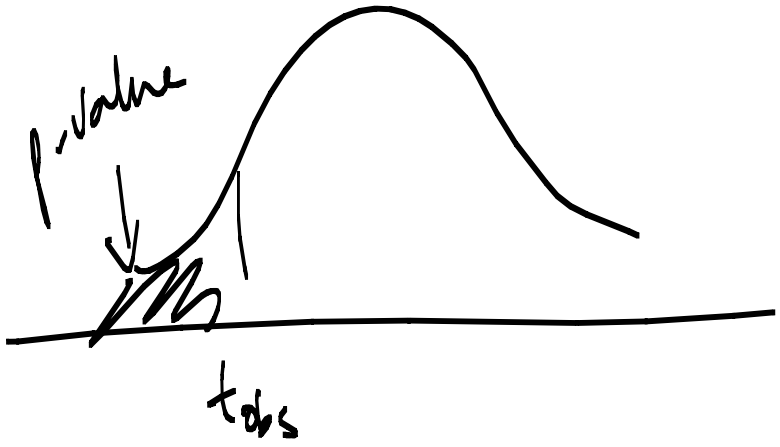
Practical conclusion: strong evidence that CFC concentration is changing over time

Suppose want to test $H_0: \beta_1 = 0$ vs $H_a: \beta_1 > 0$

test statistic calculation same



If want to test $H_0: \beta_1 = 0$ vs $H_a: \beta_1 < 0$



But only do 1-sided tests if you have a really good a priori reason.

If want to test $H_0 : \beta_1 = \beta_1^*$

test statistic $t_{obs} = \dfrac{b_1 - \beta_1^*}{S/\sqrt{S_{xx}}}$ (assume H_0 is true)

Can construct similar tests for β_0

Regression Analysis of Variance

How well does the regression line summarize the data?

Decomposition of Sums of Squares

$$y_i = \hat{y}_i + \hat{e}_i$$

$$= b_0 + b_1 x_i + \hat{e}_i$$

$$= \bar{y} - b_1 \bar{x} + b_1 x_i + \hat{e}_i$$

$$y_i - \bar{y} = b_1 (x_i - \bar{x}) + \hat{e}_i$$

Square both sides: $(y_i - \bar{y})^2 = b_1^2 (x_i - \bar{x})^2 + 2\hat{e}_i b_1 (x_i - \bar{x}) + \hat{e}_i^2$

Sum:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = b_1^2 \sum (x_i - \bar{x})^2 + \sum \hat{e}_i^2$$

$$\sum \hat{e}_i (x_i - \bar{x}) = 0$$

because

$$\sum \hat{e}_i x_i = 0$$

$$\sum \hat{e}_i = 0$$

"Analysis of Variance"

"Decomposition of SS"

SS = Sum of Squares

$\sum_{i=1}^{n} (y_i - \bar{y})^2$ = Total SS = SST

(Corrected SS)

(Uncorrected SS $= \sum y_i^2$)

$\sum_{i=1}^{n} \hat{e}_i^2$ = Residual or Error SS = RSS

- method of least squares minimized this

The term $b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$ is called

model SS or regression SS = SSReg

It is the amount of variation in y's explained by regression line.

Exercise Show $b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum (\hat{y}_i - \bar{y})^2$

Usually summarized in ANOVA table:

Source	SS	df	MS = SS/df
regression line	$b_1^2 S_{xx} = \sum (\hat{y}_i - \bar{y})^2$	1	$b_1^2 S_{xx}/1$
error	$\sum \hat{e}_i^2$	$n-2$	$\sum \hat{e}_i^2/(n-2) = s^2$
Total	$\sum (y_i - \bar{y})^2$	$n-1$	

leave blank since doesn't total anymore

Coefficient of Determination

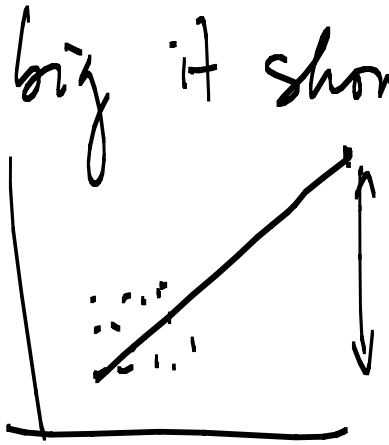
$$R^2 = \frac{SSReg}{SST} = 1 - \frac{RSS}{SST}, \quad 0 \le R^2 \le 1$$

R^2 gives percent of variation in y's that is explained by regression line

Pre-MP CFCs data $R^2 = \frac{203119}{203993} = 99.57\%$

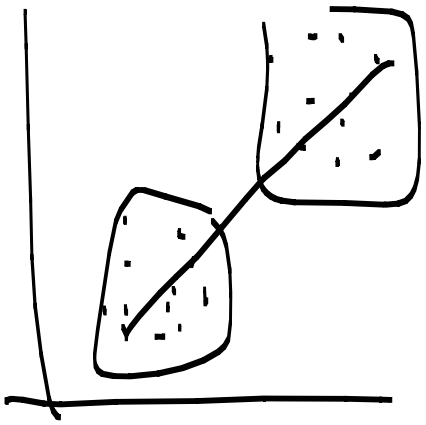
R^2 is useful <u>BUT</u>:

- no absolute rules about how big it should be
- not resistant to outliers



- not meaningful for models with no intercept

- to compare 2 models, R^2 is only useful if:
  1. same y (same observations, no transformations)
  2. one set of predictor variables is subset of other



- can get very high R^2 by overfitting

(complicated model, may fit well for data you have but won't work well on other data)

Some more distribution theory

If $U \sim \text{chisquare}(\nu_1)$, $V \sim \text{chisquare}(\nu_2)$

U, V independent

then $$\frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1, \nu_2}$$

Treat x's as non-random

Mean Square of Regression = MSReg = SSReg/1 = $b_1^2 \sum (x_i - \bar{x})^2$

Think of MSReg as estimator $\hat{\beta}_1^2 \sum (x_i - \bar{x})^2$

$E(MSReg) = \sigma^2 + \beta_1^2 S_{xx}$

$E(\hat{\beta}_1^2 S_{xx})$
$= S_{xx} E(\hat{\beta}_1^2)$
$= S_{xx} [Var(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2]$
$= S_{xx} [\sigma^2/S_{xx} + \beta_1^2]$

MSE "Mean Square Error"
$= RSS/n-2 = \sum \hat{\epsilon}_i^2 / n-2$

$E(MSE) = \sigma^2$

If $\beta_1 = 0$, $E(MSReg) = E(MSE)$

Moreover, If $\beta_1 = 0$, $\frac{MSReg}{\sigma^2} \sim$ chisquare (1)

and $\frac{MSE\,(n-2)}{\sigma^2} \sim$ chisquare $(n-2)$

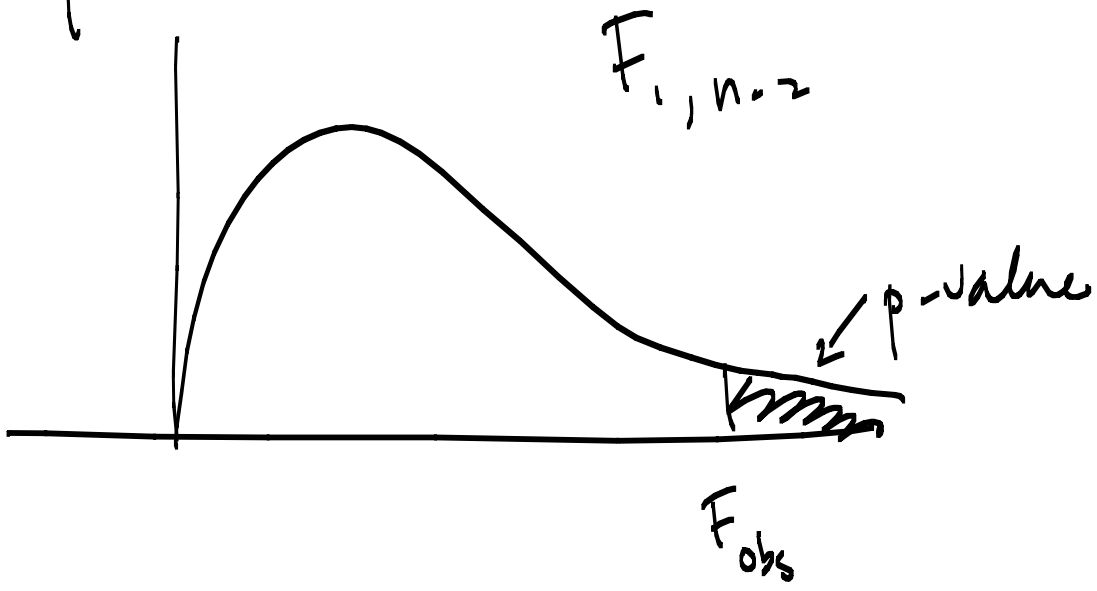
If $\beta_1 = 0$, $\frac{MSReg/\sigma^2/1}{MSE\,(n-2)/\sigma^2/n-2} \sim F_{1,n-2}$

Another test of $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Use as test statistic $F_{obs} = \frac{MSReg}{MSE}$

Under H_0, this is an observation from a F distribution with 1 and $n-2$ degrees of freedom

$\beta_1 \neq 0$ gives larger values of F_{obs}, so deviations from $\beta_1 = 0$ are in the right tail of F-distribution



Pre-MP CFCs data $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

$F_{obs} = 20319 / 5.788 = 3509 2.6$

$p < .0001$

Strong evidence that slope is not 0

Note $t_{obs}^2 = (187.33)^2 = 35092.6$

Have 2 tests for $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$

Show, in general, $(t_{obs})^2 = F_{obs}$

In general, the square of a r.v. with a t_m distribution results in a r.v. with a $F_{1,m}$ distribution.