Simple Linear Regression-II

CIVL 7012/8012

Recap(1)

- On simple linear regression, we learned
  - estimation methods
  - Gauss-Markov theorem
  - Goodness of fit
  - Interpretation
  - Incorporating non-linearities (log and exponential forms)

Significance of parameters

- The *t-test*
- t statistic for $\hat{\beta}_j$:

$$t_{\hat{\beta}_j} \equiv \hat{\beta}_j / se(\hat{\beta}_j)$$

- Null hypothesis

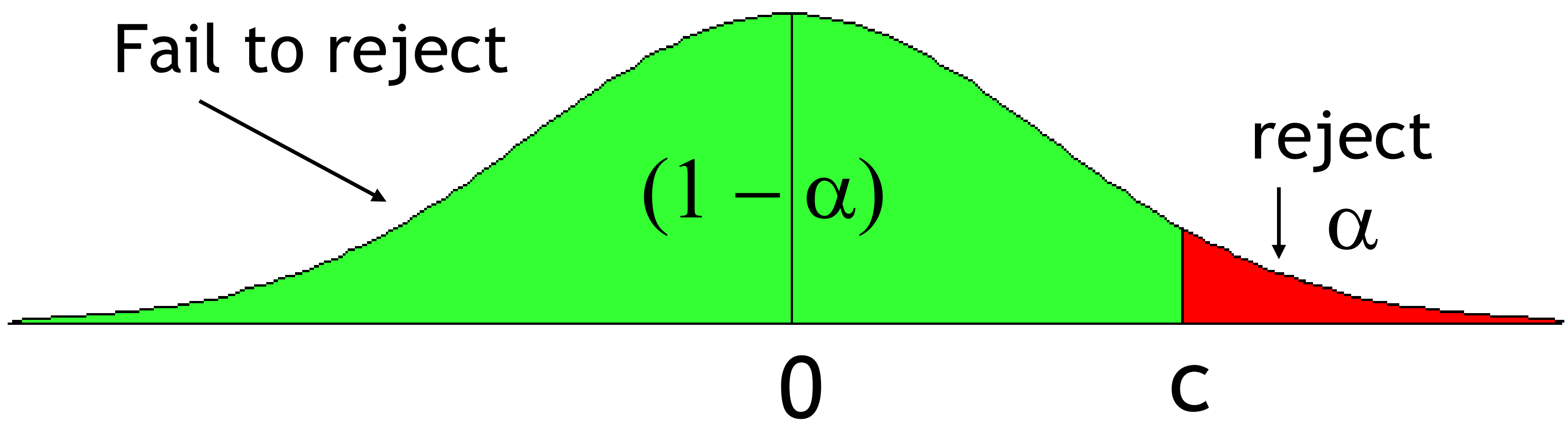
$$H_0\text{: } \beta_j = 0$$

- Alternate hypothesis
  - H_1: $\beta_j > 0$ and H_1: $\beta_j < 0$ are one-sided
  - H_1: $\beta_j \neq 0$ is a two-sided

One-sided alternative

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

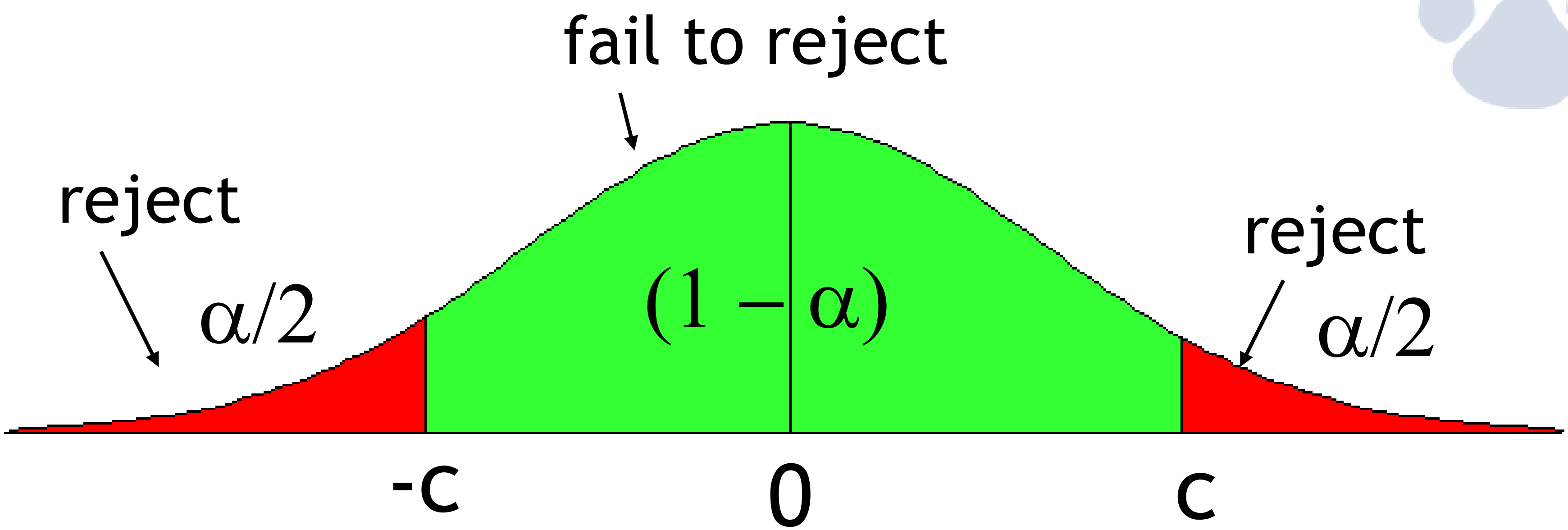
H_0: $\beta_j = 0$ H_1: $\beta_j > 0$



Two-sided alternative

$$y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + u_i$$

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



Confidence interval of parameter estimate

- Confidence interval using the same critical value as was used for a two-sided test
- A (1 - α) % confidence interval is defined as

$$\hat{\beta}_j \pm c \bullet se\left(\hat{\beta}_j\right), \text{ where } c \text{ is the } \left(1 - \frac{\alpha}{2}\right) \text{ percentile}$$

in a t_{n-k-1} distribution

Computing p-value for t-tests

- Question:
  - "what is the smallest significance level at which the null would be rejected?"
- Compute the *t* statistic, and then look up what percentile it is in the appropriate *t* distribution – this is the *p*-value
- Example
  - If p-value is less than 0.05 then the parameter is significant at 95% level of confidence

Confidence interval of mean response

A $100(1-\alpha)\%$ confidence interval on the mean response at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\hat{\mu}_{Y|x_0} - t\alpha/2,\, n-2\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]}$$

$$\le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]} \qquad (11\text{-}31)$$

where $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is computed from the fitted regression model.

Prediction interval of new response

A $100(1-\alpha)\%$ **prediction interval on a future observation** Y_0 at the value x_0 is given by

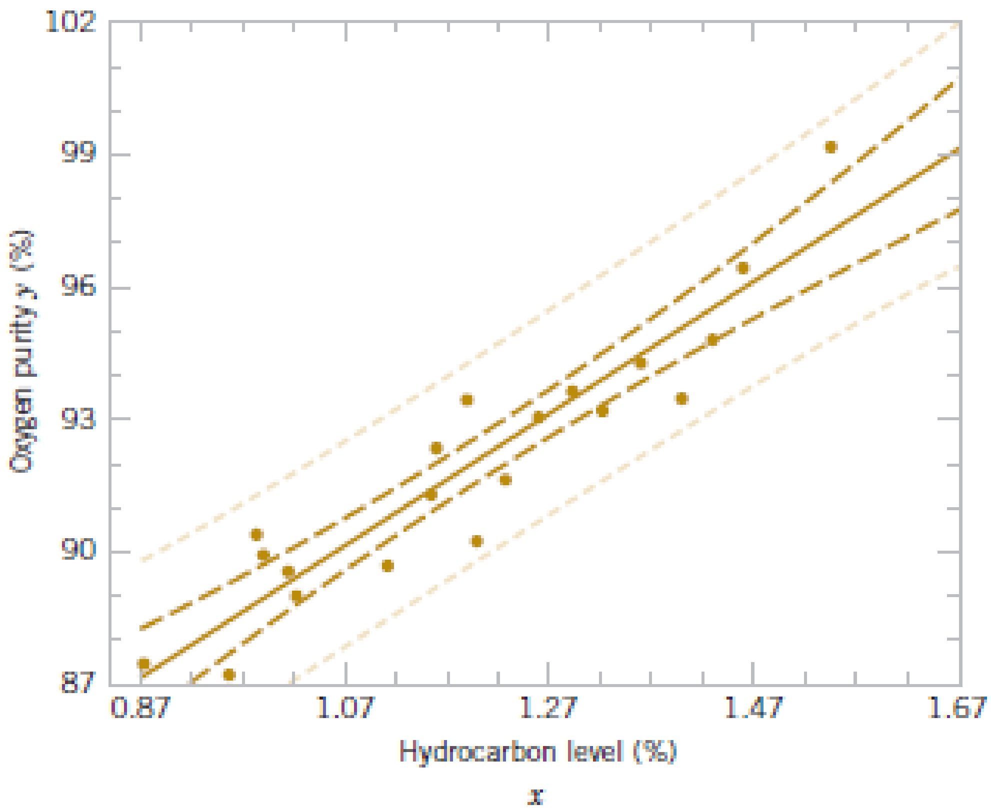
$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]}$$

$$\le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}\right]} \qquad (11\text{-}33)$$

The value $\hat{y}_0$ is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Prediction and confidence interval

- Prediction interval is larger than confidence interval



Regression Passing through Origin

- Regression equation becomes: $\tilde{y} = \tilde{\beta}_1 x,$
- Using OLS: $\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_i)^2.$
- First order conditions: $\sum_{i=1}^{n} x_i(y_i - \tilde{\beta}_1 x_i) = 0.$
- Parameter estimate: $\tilde{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2},$

Regression Passing through Origin

- R-square becomes

$$1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

- This term can be negative
- Means, using simple averages to predict y is better than using regression equation passing through origin

Regression of a constant

- No need to have x (no variability)
- Intercept itself is mean of y
- No parameter estimates needed