TDAB01 Probability and Statistics

Maryna Prus
IDA, Linköping University

Lecture 12: Prediction

Overview

- **ANOVA and R^2**
- **Inference about regression slope**
- **Confidence interval for mean response**
- **Prediction interval for individual response**

ANOVA and R^2

- Regression: Predict $E[Y|X = x]$ where Y is a random variable (response variable or dependent variable) and $X = x$ is an observation (explanatory variable or independent variable)
- Linear Regression: $E[Y|X = x] = \mu(x) = \beta_0 + \beta_1 x$ where
  - β_0 is intercept
  - β_1 is slope
- Least squares or maximum likelihood estimators for β_0 and β_1:
  - $b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and
  - $b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$
- $SS_{TOTAL} = \sum_i (y_i - \bar{y})^2$ = the **total** variation of Y
- $SS_{REG} = \sum_i (\hat{y}_i - \bar{y})^2$ = that variation **explained** of the model
- $SS_{ERR} = \sum_i (y_i - \hat{y}_i)^2$ = that variation **not** explained by the model
  $SS_{ERR} = SS_{TOTAL} - SS_{REG}$
- $R^2 = \frac{SS_{REG}}{SS_{TOTAL}}$ = **proportion** of the total variation explained by the model
- $0 \le R^2 \le 1$
- For linear regression $R^2 = r^2$ - squared sampling correlation coefficient between X and Y

Inference about regression slope

- Note that: $\sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = 0$ and then

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x}) y_i$$

- $b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x}) y_i}{S_{xx}}$ and then b_1 is a linear function of y_i and is hence **normally** distributed (if Y_i is normal)
- $E[b_1] = \frac{\sum_i (x_i - \bar{x}) E[y_i]}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} = \frac{\beta_1 \sum_i (x_i - \bar{x}) x_i}{S_{xx}} = \beta_1$ and then b_1 is a **unbiased** estimator of β_1.
- $var[b_1] = \frac{\sum_i (x_i - \bar{x})^2 var[y_i]}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$.
- Now we can build confidence intervals and hypothesis tests for the slope based on the t distribution, since σ^2 is usually unknown.
  - $(1 - \alpha)100\%$ two-sided confidence interval:

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}$$

    where the t distribution has $n - 2$ degrees of freedom, and $s^2 = SS_{ERR}/(n-2)$.
  - Hypothesis test $H_0 : \beta_1 = B$ vs $H_A : \beta_1 \neq B$:

$$t = \frac{b_1 - B}{s/\sqrt{S_{xx}}}$$

    which has a t distribution has $n - 2$ degrees of freedom. Take $B = 0$ to test if there is a linear relationship between X and Y.

Confidence interval for mean response

- $\mu_* = \mu(x_*) = E[Y|X = x_*] = \beta_0 + \beta_1 x_*$ is estimated by $\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = \bar{y} - b_1\bar{x} + b_1 x_* = \bar{y} + b_1(x_* - \bar{x}) = \frac{\sum_i y_i}{n} + \frac{\sum_i (x_i - \bar{x}) y_i (x_* - \bar{x})}{S_{xx}} = \sum_i \left(\frac{1}{n} + \frac{\sum_i (x_i - \bar{x})(x_* - \bar{x})}{S_{xx}} \right) y_i$ and then $\hat{y}_*$ is a linear function of y_i and then **normal** distributed.
- Note. that we predict a **population parameter**.
- $E[\hat{y}_*] = E[b_0] + E[b_1]x_* = \beta_0 + \beta_1 x_* = \mu_*$ and then $\hat{y}_*$ is a **expectation correct** estimator of μ_*.
- $var[\hat{y}_*] = \sum_i \left(\frac{1}{n} + \frac{\sum_i (x_i - \bar{x})(x_* - \bar{x})}{S_{xx}} \right)^2 var(y_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right)$.
- $(1 - \alpha)100\%$ two-sided confidence interval:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

  where the t distribution has $n - 2$ degrees of freedom, and $s^2 = SS_{ERR}/(n - 2)$.

Prediction interval for individual response

- A confidence interval for $\hat{y}_*$ represents uncertainty about **population expectation** at $X = x_*$. But how does the uncertainty of **random variable** Y value look like if $X = x_*$?
- 95% **prediction interval** for the Y value is an interval $[a, b]$ such that

$$\boldsymbol{P}(a \le Y \le b | X = x_*) = 0.95$$

  where a, b and Y are random variables
- $(1 - \alpha)100\%$ prediction interval for Y given $X = x_*$:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

  where the t distribution has $n - 2$ degrees of freedom, and $s^2 = SS_{ERR}/(n - 2)$.

Prediction interval for individual response

- $(1-\alpha)100\%$ confidence interval:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

- $(1-\alpha)100\%$ prediction interval:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

- Length of prediction interval is larger than length of confidence interval, i.e. predicting an individual response variable is more difficult than predicting the population's expected value.
- Example: See Example 11.7 in textbook

Thank you for your attention!