Biost 517 / Biost 514
Applied Biostatistics II / Biostatistics II

Timothy A. Thornton, Ph.D.
Associate Professor of Biostatistics
University of Washington

Lecture 15:

Theory of Classical Linear Regression

Least-Squares Regression

- Consider a sample of size n from a population, and let Y_i and X_i be the response and predictor variables, respectively, for individual i in the sample
- In the previous lecture, we noted that simple linear regression is often referred to as "Least-Squares regression" because the linear regression line has an intercept and slope that minimizes the total squared distance of each response to the line, i.e., the β_0 and β_1 that minimizes the function:

$$\sum_{i=1}^{n}\left[Y_i-\left(\beta_0+\beta_1 X_i\right)\right]^2$$

- The least squares line can be obtained with multivariate calculus

Statistical Inference of Regression Parameters

- We, however, are interested in obtaining "statistical inference" about a population parameter θ from a linear regression line that is calculated using data on a random sample from the population
- For appropriate statistical inference, we must take into account that the parameter is **estimated** from the sample
- There will be error in any estimates of a parameter with the linear regression line due to sampling variability of the response (as well as the predictor of interest).
- For statistical inference of the population parameter, we need, at a minimum:
  - an estimate of the parameter
  - standard errors of the estimate (for confidence intervals, hypothesis testing, p-values, etc.)

Classical Linear Regression

- In classical linear regression, for each individual i the response variable conditional on the predictor is assumed to have the following normal distribution: $Y_i \mid X_i \sim N(\theta_i, \sigma^2)$ where

$$\theta_i = E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

- So

$$f(Y_i \mid X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \left(\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma^2} \right)^2 \right]$$

- For a sample with n independent individuals, the likelihood is:

$$\prod_i^n f(Y_i \mid X_i)$$

- The maximum likelihood estimates (MLE) for β_0 and β_1 correspond to the intercept and slope for the least-squares regression line! Why?

Maximum Likelihood Estimates

- The maximum likelihood estimate (MLE) for β_0 and β_1 are

$$\widehat{\beta}_1 = r_{XY} \frac{SD_Y}{SD_X}$$

and

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

  - r_{XY} is the sample correlation between X and Y in the sample
  - SD_Y is the standard deviation of Y
  - SD_X is the standard deviation of X

Fitted Values

- So, the MLE of $E(Y_i \mid X_i)$ is

$$\widehat{E[Y_i \mid X_i]} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- $\hat{Y}_i$ is often referred to as the “fitted value"
- We now need to obtain standard errors of the two estimates of the parameters.

Residuals

- The residual for individual i in a regression is defined to be the observed value minus the predicted (or fitted) value:

$$e_i = Y_i - \hat{Y}_i$$

- And the residual sum of squares is $RSS = \sum_{i=1}^{n} \left(e_i\right)^2$

- The $Var(Y_i) = \sigma^2$ can be estimated by

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

- Why is this a reasonable estimate for the variance of Y?

Estimated Variances

- The estimated variances of the regression coefficients are

$$\widehat{Var}(\hat{\beta}_1) = \hat{\sigma}^2 \frac{1}{(n-1)SD_X^2}$$

and

$$\widehat{Var}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)SD_X^2} \right)$$

Confidence Intervals of Regression Parameters

- Using statistical theory, it can be shown that the parameter estimates for the linear regression mean model will be normally distributed because they are linear combinations of the Y_i's
- Confidence intervals and tests statistics of the parameters can now be obtained using familiar methods based on the **t distribution:**

$(1-\alpha)*100\%$ confidence interval can be obtained by:

$$\hat{\beta}_j \pm t(\alpha, n-2) \times SE(\hat{\beta}_j)$$

for $j \in \{0,1\}$, where $SE(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)}$ and $t(\alpha, n-2)$ is the critical value of a t distribution with $n-2$ degree of freedom, i.e., the value such that the area to the right of this value under the density curve for the t distribution is $\alpha / 2$

Hypothesis Testing of Regression Parameters

- A hypothesis test of $H_0 : \beta_j = \beta_j^*$ versus $H_a : \beta_j \neq \beta_j^*$ can be obtained by computing the t test statistics:

$$t_{stat} = \frac{\hat{\beta}_j - \beta_j^*}{SE(\hat{\beta}_j)}$$

- Under the null hypothesis, the test statistic follows a t distribution with $n-2$ degrees of freedom under

Interpreting Regression Parameters

$$\widehat{E[Y_i \mid X_i]} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- The intercept $\hat{\beta}_0$ is the estimated mean value of Y_i for a group with $X_i = 0$
  - Quite often not of scientific interest
  - Often outside range of data, sometimes not even scientifically possible

Interpreting Regression Parameters

$$\widehat{E[Y_i \mid X_i]} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- The slope of the linear regression line is

$$\hat{\beta}_1 = r_{XY} \frac{SD_Y}{SD_X}$$

- A change of 1 SD in X corresponds to a change of r_{XY} SDs in Y.
- Slope also corresponds to the difference in mean Y across groups differing in X by 1 unit
    - Usually measures association between Y and X

$(\bar{X}, \bar{Y})$

Derivation of Interpretation

- Simple linear regression of response Y on predictor X
  - Mean for an arbitrary group derived from model
  - Interpretation of parameters by considering special cases

Model	$E[Y_i \mid X_i] = \beta_0 + \beta_1 \times X_i$
$X_i = 0$	$E[Y_i \mid X_i = 0] = \beta_0$
$X_i = x$	$E[Y_i \mid X_i = x] = \beta_0 + \beta_1 \times x$
$X_i = x+1$	$E[Y_i \mid X_i = x+1] = \beta_0 + \beta_1 \times x + \beta_1$

Equal Variance Assumption

- The classical linear regression model assumes that the distribution of the response is normal for each subpopulation grouping defined by the POI, and equal variance for all subpopulations.
- We will soon discuss “robust” regression methods that allow for the variances to be different across groups.