Lecture 2
Inference for Simple Linear Regression
STAT 632, Spring 2020

Assumptions for SLR

1. **Linearity**: Y is related to x by a simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$ with mean $E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$. That is, the data follow a linear trend in the scatter plot between X and Y.
2. **Independence**: The errors $e_1, e_2, \cdots, e_n$ are independent of each other.
3. **Constant Variance**: The errors $e_1, e_2, \cdots, e_n$ have common variance $Var(e_i) = \sigma^2$.
4. **Normality**: The errors are normally distributed, i.e, $e_i \sim N(0, \sigma^2)$

Remark: The assumptions are necessary for making inferences about the least squares estimates for the slope and intercept (i.e., hypothesis testing and confidence intervals), and for constructing valid prediction intervals.

Simple linear regression model for the population:

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

β_0 and β_1 are the population parameters (fixed and non-random)

Least squares line (estimated from the sample):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates (random, varies from sample to sample)

Inferences About the Slope

Recall the least squares estimate of $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

Since, $\sum_{i=1}^n (x_i - \bar{x}) = 0$ we find that

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n (x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})y_i$$

Thus, we can rewrite $\hat{\beta}_1$ as $\hat{\beta}_1 = \sum_{i=1}^n c_i y_i$, where $c_i = \frac{x_i - \bar{x}}{\mathrm{SXX}}$

Inferences About the Slope

Under the assumptions for SLR, the expectation and variance of the least squares estimate of the slope is given by

$$E(\hat{\beta}_1) = \beta_1$$
$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\text{SXX}}$$

Thus, the sampling distribution for $\hat{\beta}_1$ is

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\text{SXX}})$$

Derivation – use result from previous slide.

Inferences About the Slope

Standardizing gives

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{\text{SXX}}} \sim N(0,1)$$

However, since σ is unknown we replace it with $\hat{\sigma}$, giving

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{\text{SXX}}} = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)},$$

which follows a t distribution with $n-2$ degrees of freedom (sample size - number of parameters estimated).

Inferences About the Slope

Test whether the slope β_1 is zero. That is, test whether or not there is a linear association between X and Y.

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

Test statistic:

$$T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}; \quad \text{df=n-2}$$

$1-\alpha$ confidence interval for the slope β_1:

$$\hat{\beta}_1 \pm t_{\alpha/2;n-2}se(\hat{\beta}_1)$$

Inferences About the Intercept

Recall the least squares estimate of the intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Under the assumptions for SLR, the expectation and variance is given by

$$E(\hat{\beta}_0) = \beta_0$$
$$Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}\right)$$

Thus, the sampling distribution for $\hat{\beta}_0$ is

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}\right)\right)$$

Derivation provided in Sheather, section 2.7.2

Inferences About the Intercept

Standardizing gives,

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} \sim N(0,1)$$

However, since σ is unknown we replace it with $\hat{\sigma}$, giving

$$T = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}} = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)},$$

which follows a t distribution with $n-2$ degrees of freedom.

Inferences About the Intercept

Test whether the intercept β_0 is zero.

$H_0 : \beta_0 = 0$
$H_A : \beta_0 \neq 0$

Test statistic:

$$T = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}; \quad \text{df=n-2}$$

$1-\alpha$ confidence interval for the intercept β_0:

$$\hat{\beta}_0 \pm t_{\alpha/2;n-2} se(\hat{\beta}_0)$$

Example

```
> lm1 <- lm(wgt ~ hgt, data=bdims_males)
> summary(lm1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -60.95336   14.05436  -4.337 2.11e-05 ***
hgt           0.78257    0.07901   9.905  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.902 on 245 degrees of freedom
Multiple R-squared:  0.2859,Adjusted R-squared:  0.283
F-statistic: 98.11 on 1 and 245 DF,  p-value: < 2.2e-16
```

Example

```
> confint(lm1)
                  2.5 %     97.5 %
(Intercept) -88.6361527 -33.270576
hgt           0.6269509   0.938186

# manual calculation CI for slope
> n <- nrow(bdims_males)
> tcrit <- qt(0.975, df=n-2)
> 0.78257 - tcrit * 0.07901
[1] 0.6269445
> 0.78257 + tcrit * 0.07901
[1] 0.9381955
```