Inference in Linear Regression

There are four basic assumptions made about the relationship between a response variable Y and an explanatory variable X in simple linear regression.

1. All Y values are independent of one another.
2. For each value of X, the distribution of possible Y values is normal.
3. The normal distribution for Y values corresponding to a particular value of X has a mean $\mu\{Y|X\}$ that lies on the straight line
$$\mu\{Y|X\} = \beta_0 + \beta_1 X.$$
This line is called the population regression line. The parameter β_0 is the intercept of the population regression line. It represents the mean of the Y values when $X = 0$. The parameter β_1 is the slope of the population regression line. It represents the change in the mean of Y per unit increase in X.
4. The normal distribution of Y values corresponding to a particular value of X has standard deviation $\sigma\{Y|X\}$. That standard deviation is usually assumed to be the same for all values of X so that we may write $\sigma\{Y|X\} = \sigma$.

Suppose we have n observations of a response variable Y and an explanatory variable X: $(X_1, Y_1), \ldots, (X_n, Y_n)$.

The simple linear regression model can be written succinctly as $Y_i = \beta_0 + \beta_1 X_i + e_i$ for $i = 1, \ldots, n$.

Here $e_1, \ldots, e_n$ are assumed to be independent normal random variables (often called random errors) with mean 0 and standard deviation $\sigma\{Y|X\} = \sigma$.

We do not get to see $e_1, \ldots, e_n$; but we can approximate them by the residuals $\hat{e}_1, \ldots, \hat{e}_n$.

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \quad \text{approximates} \quad Y_i - (\beta_0 + \beta_1 X_i) = e_i.$$

An estimate of σ is given by $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^n \hat{e}_i^2}{n-2}}$. Note that this is the square root of the residual sum of squares divided by its degrees of freedom. We also used the square root of the residual sum of squares to estimate σ in one-way ANOVA.

An alternative expression for $\hat{\sigma}$ is $s_Y \sqrt{(1 - r^2)\frac{n-1}{n-2}}$.

The population parameter β_1 is estimated by the slope of the least-squares regression line $\hat{\beta}_1$.

$$\text{Mean}(\hat{\beta}_1) = \beta_1 \qquad \text{SE}(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)s_X^2}} \qquad t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \qquad d.f. = n - 2$$

The population parameter β_0 is estimated by the intercept of the least-squares regression line $\hat{\beta}_0$.

$$\text{Mean}(\hat{\beta}_0) = \beta_0 \qquad \text{SE}(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}} \qquad t = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \qquad d.f. = n - 2$$

Using data from 56 U.S. cities collected from 1931 to 1960, we computed the equation of the least-squares regression line relating average minimum January temperature (Y) to latitude in degrees north of the equator (X) as $\hat{Y} = 108.56 - 2.104X$. To determine the equation of the least-squares regression line we used the following summary statistics:

$$\bar{X} = 39.0 \qquad \bar{Y} = 26.5 \qquad S_X = 5.4 \qquad S_Y = 13.4 \qquad r = -0.848.$$

Use this information to help you complete the following problems.

1. Estimate β_0.

2. Estimate β_1.

3. Compute a 95% confidence interval for β_1.

4. Is there a significant linear relationship between average minimum January temperature and latitude? Test $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$.

5. Which of the four assumptions of simple linear regression can you check by examining a scatter plot of Y vs. X? Do any of the assumptions that you can check appear to be violated?