11.8 Correlation Model

- In the regression context, we estimate $E(Y)$ and predict Y based on fixed values of X.
- In some cases, we view both X and Y as random variables and seek only to study the <u>association</u> between X and Y.
- Recall from STAT 511 that the correlation coefficient ρ measures the <u>linear</u> association between X and Y.
- If $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid from a bivariate normal distribution, then the MLE of ρ is the <u>sample correlation coefficient</u>:

Since the SLR slope estimate
$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$,

- If $(X, Y) \sim$ bivariate normal and if the linear regression model holds,

- Tests of $H_0: \rho = 0$ (which under bivariate normality correspond to testing whether X and Y are independent) are equivalent to our t-tests for $H_0: \beta_1 = 0$ in SLR.
- The test statistic is

which is algebraically equivalent to the SLR test statistic

- This test statistic has a ____________ with ____________ under H_0.
- Therefore the usual t-test RR can be used to test $H_0: \rho = 0$ against whichever appropriate alternative.
- To test $H_0: \rho = \rho_0$ for some nonzero ρ_0, we must use <u>Fisher's z-transformation.</u>
- For large samples,

So we use the test statistic

- The RR for an α-level test is:

- A $100(1-\alpha)\%$ large-sample CI $[L, U]$ for can be obtained via:

- The $100(1-\alpha)\%$ CI for ρ can then be found by back-transforming the endpoints L and U:

<u>Example 1</u>: The weights (X, in thousands of pounds) and gas mileages (Y, in mpg) for 32 randomly selected cars were collected. Assume bivariate normality for (X,Y). Summary calculations yield $S_{xy} = -158.617$, $S_{xx} = 29.679$, $S_{yy} = 1126.047$. Estimate ρ and test (at $\alpha = .05$) whether X and Y are independent.

<u>Example 1(a)</u>: Test (at $\alpha = .05$) whether the correlation coefficient between X and Y is less than -0.7. Find a 95% CI for ρ.

- The cor.test function in R can perform these calculations easily.

The Coefficient of Determination

- Note that $S_{yy} = \sum (y_i - \bar{y})^2$ is a measure of the sample variation in the y-values.
- Recall that $SSE = \sum (y_i - \hat{y}_i)^2$ measures the amount of variation in the y's unexplained by the linear regression model.
- In the context of SLR,

is called the coefficient of determination and measures the proportion of variation in the y-values explained by the linear model.

Example 1: $r =$

So

Note: In MLR, a coefficient of multiple determination is defined similarly:

and has a similar interpretation.