STAT 518 --- Section 5.5: Distribution-Free Tests in Regression

- Suppose we gather data on two random variables.
- We wish to determine: Is there a relationship between the two r.v.'s? (correlation and/or regression)
- Can we use the values of one r.v. (say, X) to predict the other r.v. (say, Y)? (regression) ← independent variable
- Often we assume a straight-line relationship between two variables. ← dependent (response) variables.
- This is known as <u>simple linear regression</u>.

<u>Example 1</u>: We want to predict Y = breathalyzer reading based on X = amount of alcohol consumed.
<u>Example 2</u>: We want to estimate the effect of a medication dosage on the blood pressure of a patient.
<u>Example 3</u>: We want to predict a college applicant's college GPA based on his/her SAT score.

- This again assumes we have <u>paired</u> data (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n) for the two related variables.

Linear Regression Model

- The linear regression model assumes that the mean of Y (for a specific value x of X) varies linearly with x:

$$E[Y \mid X = x] = \alpha + \beta x$$

α = intercept and β = slope

• These parameters are **unknown** and must be **estimated** using sample data.

• Estimating the unknown parameters is also called **fitting the regression model**.

Fitting the Model (Least Squares Method)

• If we gather data (X_i, Y_i) for several individuals, we can use these data to estimate α and β and thus estimate the linear relationship between Y and X.

• Once we settle on the "best-fitting" regression line, its equation gives a predicted Y-value for any new X-value:

$$\hat{Y} = a + bX$$

• How do we decide, given a data set, which values a and b produce the best-fitting line?

• For each point, the **error** = $D_i = Y_i - (a + bX_i)$

(Some positive errors, some negative errors)

• We want the line that makes these errors as small as possible (so that the line is "close" to the points).

Least-squares method: We choose the line that minimizes the sum of all the **squared** errors (SSE).

want to minimize $\sum_{i=1}^{n} D_i^2$

Least squares estimates a and b:

$$b = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}, \quad a = \bar{Y} - b\bar{X}$$

- **This least-squares method is completely distribution-free.**

- **In classical models, we must assume** normality **of the data in order to perform parametric inference.**

- **Since the slope β describes the marginal effect of *X* on *Y*, we are most often interested in hypothesis tests and confidence intervals about β.**

- **If the data are normal, these are based on the *t*-distribution.**

- **If the data's distribution is unknown, we can use a nonparametric approach.**
- **We must assume only that the *Y*'s are independent, identically distributed, and that the *Y*'s and *X*'s are at least interval in measurement scale.**
- **We further assume that the residual** $Y - E(Y|X)$ is independent of X.

A Distribution-Free Test about the Slope

- **Let β_0 be some hypothesized value for the slope.**

- **For each bivariate observation, compute**

$$U_i = Y_i - \beta_0 X_i$$

and calculate the Spearman's rho for the pairs

$$(X_1, U_1), \dots, (X_n, U_n)$$

If $H_0: \beta = \beta_0$ is true, then the X_i's and the U_i's are independent.

Do Spearman test of independence on the X_i's and U_i's.

Hypotheses and Decision Rules

Table A10

Two-tailed	Lower-tailed	Upper-tailed
$H_0: \beta = \beta_0$	$H_0: \beta \geq \beta_0$	$H_0: \beta \leq \beta_0$
$H_1: \beta \neq \beta_0$	$H_1: \beta < \beta_0$	$H_1: \beta > \beta_0$
Reject H_0 if $\lvert\rho\rvert > w_{1-\alpha/2}$	Reject H_0 if $\rho < -w_{1-\alpha}$	Reject H_0 if $\rho > w_{1-\alpha}$

A Distribution-Free Confidence Interval for the Slope

- **For each pair of points** (X_i, Y_i) and (X_j, Y_j), $i<j$ and $X_i \neq X_j$

compute the "two-point slope":

$$S_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

- **There are, say, *N* such "two-point slopes".**

- **Let the ordered two-point slopes be:**

$$S^{(1)} \leq S^{(2)} \leq \cdots \leq S^{(N)}$$

- **For a $(1-\alpha)100\%$ CI, find $w_{1-\alpha/2}$ from Table A11 and define *r* and *s* as:**

$$r = \frac{1}{2}\left(N - w_{1-\alpha/2}\right)$$

$$s = \frac{1}{2}\left(N + w_{1-\alpha/2}\right) + 1 = N + 1 - r$$

- **If *r* and *s* are not integers, round *r* down to the next smallest integer and round *s* up to the next largest integer (in order to produce a conservative CI).**

- **The $(1-\alpha)100\%$ CI for β is then**

$$\left(S^{(r)},\ S^{(s)}\right)$$

- **This CI will have coverage probability of <u>at least</u> $1-\alpha$.**

Example 1 (GMAT/GPA data): Recall example from Section 5.4. Suppose a national study reports that an increase of 40 points in GMAT score yields a 0.4 expected increase in GPA. Does this sample provide evidence against that claim? (Use $\alpha = 0.05$.)

$Y = GPA \quad X = GMAT \qquad \frac{0.4}{40} = 0.01$

Test $H_0: \beta = 0.01$ vs. $H_1: \beta \neq 0.01$

Here, $U_i = Y_i - 0.01 X_i$

From R, Spearman's ρ for X_i's and U_i's is: -0.728

$w_{.975} = .5804 \leftarrow$ from Table A10.
for $n=12$

$|\rho| = .728 > .5804$, so we reject H_0.

- Conclude the true slope is <u>not</u> 0.01.
- An increase in GMAT of 40 points does <u>not</u> yield an increase in expected GPA of 0.4 points.
- From R, P-value $= .0072$.

95% CI for β: $(0.000, 0.008)$ from R. ← using $w_{.975} = 28$ from Table A11.

(This is conservative – has coverage probability <u>at</u> <u>least</u> 0.95)

• **In cases with severe outliers, the least-squares estimated slope can be severely affected by such outliers. An alternative set of regression estimates was suggested by Theil:**

b_1 = sample median of S_{ij}'s

a_1 = median(Y_i's) $-$ b_1 [median(X_i's)]

$\Rightarrow \hat{Y} = a_1 + b_1 X$

Example 2: For several levels of drug dosage (*X*), a lipid measure (*Y*) is taken. The data are:

X:	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Y:	**2.5**	**3.1**	**3.4**	**4.0**	**4.6**	**11.1**	**5.1**

• **See R code for example plots using the least-squares line and Theil's regression line.**

• **The point estimator of the slope in Theil's method is called the Hodges-Lehmann estimator.**

Comparison to Competing Tests

• **When the distribution of (*X*, *Y*) is bivariate normal and the X_i's are equally spaced, the nonparametric test for the slope has A.R.E. of** 0.98 **relative to the classical t-test.**
• **In general, this A.R.E. is always at least** 0.95**.**