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)
• Often we assume a straight-line relationship between two 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:

α =   and β =

• These parameters are <u>unknown</u> and must be <u>estimated</u> using sample data.

• Estimating the unknown parameters is also called <u>fitting the regression model</u>.

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:

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

• For each point, the <u>error</u> =
(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).

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

Least squares estimates a and b:

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

• In classical models, we must assume _______________ 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

A Distribution-Free Test about the Slope

• Let β_0 be some hypothesized value for the slope.

• For each bivariate observation, compute

and calculate the Spearman's rho for the pairs

Hypotheses and Decision Rules

Two-tailed **Lower-tailed** **Upper-tailed**

A Distribution-Free Confidence Interval for the Slope

- **For each pair of points**

compute the "two-point slope":

- **There are, say, *N* such "two-point slopes".**
- **Let the ordered two-point slopes be:**
- **For a $(1 - \alpha)100\%$ CI, find $w_{1-\alpha/2}$ from Table A11 and define *r* and *s* as:**
- **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

• 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 $\alpha = 0.05$.)

• 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:

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 <u>Hodges-Lehmann estimator</u>.

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 _______ relative to the classical t-test.
• In general, this A.R.E. is <u>always</u> at least _________.

Nonparametric Regression

• Section 5.6 gives a rank-based procedure for estimating a regression function when the function is <u>unknown</u> and <u>nonlinear</u> BUT known to be <u>monotonic</u>.

• Here we will examine a distribution-free method of estimating a very general type of regression function.

• In nonparametric regression, we assume very little about the functional form of the regression function.

• We assume the model:

where $f(\cdot)$ is unknown but is typically assumed to be a smooth and continuous function.
• We also assume independence for the residuals

<u>Goal</u>: Estimate the mean response function $f(\cdot)$.

Advantages of Nonparametric Regression

• Useful when we cannot know the relationship between Y and X
• More flexible type of regression model
• Can account for unusual behavior in the data
• Less likely to have bias resulting from wrong model being chosen

Disadvantages of Nonparametric Regression

• Not as easy to interpret
• No easy way to describe relationship between *Y* and *X* with a formula (must be done with a graph)
• Inference is not as straightforward

<u>Note</u>: Nonparametric regression is sometimes called ________________ ____________.

Kernel Regression

• The idea behind kernel regression is to estimate $f(x)$ <u>at each</u> value x^* along the horizontal axis.

• At each value x^*, the estimate is simply an

• Consider a "window' of points centered at x^*:

• **The width of this window is called the _____________.**

• **At each different x^*, the window of points __________ to the left or right**

• **Better idea: Use**

• **This can be done using a ________________ function known as a <u>kernel</u>.**

• **Then, for any x^*,**

where the weights

$K(\cdot)$ is a kernel function, which typically is a <u>density</u> function symmetric about 0.

λ = bandwidth, which controls the <u>smoothness</u> of the estimate of $f(x)$.

Possible choices of kernel:

Pictures:

<u>Note</u>: The Nadaraya-Watson estimator

is a modification that assures that the weights for the Y_i's will sum to one.

• The choice of <u>bandwidth</u> λ is of more practical importance than the choice of kernel.

• The bandwidth controls how many data values are used to compute $f(x^*)$ at each x^*.

Large $\lambda \rightarrow$

Small $\lambda \rightarrow$

• Choosing λ too large results in an estimate that _______________ the true nature of the relationship between *Y* and *X*.

• Choosing λ too small results in an estimate that follows the "noise" in the data too closely.

• Often the best choice of λ is made through visual inspection (pick the roughest estimate that does not fluctuate implausibly?).

• Automatic bandwidth selection methods such as <u>cross-validation</u> are also available – this chooses the λ that minimizes a mean squared prediction error.

Example: We have data on the horsepower (*X*) and gas mileage (*Y*, in miles per gallon) of 82 cars, from Heavenrich et al. (1991).

• On computer: The R function `ksmooth` performs kernel regression (see web page for examples with various kernel functions and bandwidths).