STAT 515 -- Chapter 11: Regression

- **Mostly we have studied the behavior of a single random variable.**
- **Often, however, we gather data on two random variables.**
- **We wish to determine: Is there a relationship between the two r.v.'s?**
- **Can we use the values of one r.v. to predict the other r.v.?**
- **Often we assume a straight-line relationship between two variables.**
- **This is known as <u>simple linear regression</u>.**

Probabilistic vs. Deterministic Models

If there is an exact relationship between two (or more) variables that can be predicted with certainty, without any random error, this is known as a <u>deterministic relationship</u>.

Examples:

In statistics, we usually deal with situations having random error, so <u>exact</u> predictions are not possible.

This implies a <u>probabilistic</u> relationship between the 2 variables.

Example: Y **= breathalyzer reading**
X **= amount of alcohol consumed (fl. oz.)**

Note that usually, no line will go through all the points in the data set.

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 regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1 that produce the best-fitting line in the least squares sense.

Formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$:

Estimated slope and intercept:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \text{ and } \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

where $SS_{xy} = \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$ **and**

$$SS_{xx} = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

and *n* = the number of observations.

Example (Table 11.3):

Y **=**

X **=**

SS_{xy} =

SS_{xx} =

Interpretations:

Slope:

Intercept:

Example:

Avoid extrapolation: predicting/interpreting the regression line for X-values outside the range of *X* in the data set.

Model Assumptions

Recall model equation: $Y = \beta_0 + \beta_1 X + \varepsilon$

To perform inference about our regression line, we need to make certain assumptions about the random error component, ε. We assume:

(1) The mean of the probability distribution of ε is 0. (In the long run, the values of the random error part average zero.)

(2) The variance of the probability distribution of ε is constant for all values of *X*. We denote the variance of ε by σ^2.

(3) The probability distribution of ε is normal.

(4) The values of ε for any two observed Y-values are independent – the value of ε for one Y-value has no effect on the value of ε for another Y-value.

Picture:

Estimating σ^2

Typically the error variance σ^2 is unknown.

An unbiased estimate of σ^2 is the mean squared error (MSE), also denoted s^2 sometimes.

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

where $\text{SSE} = SS_{yy} - \hat{\beta}_1 SS_{xy}$

and $SS_{yy} = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$

Note that an estimate of σ is

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

Since ε has a normal distribution, we can say, for example, that about 95% of the observed Y-values fall within $2s$ units of the corresponding values $\hat{Y}$.

Testing the Usefulness of the Model

For the SLR model, $Y = \beta_0 + \beta_1 X + \varepsilon$ **.**

Note: ***X*** **is completely useless in helping to predict** ***Y*** **if and only if** $\beta_1 = 0$**.**

So to test the usefulness of the model for predicting ***Y*****, we test:**

If we reject H_0 **and conclude** H_a **is true, then we conclude that** ***X*** **does provide information for the prediction of** ***Y*****.**

Picture:

Recall that the estimate $\hat{\beta}_1$ is a statistic that depends on the sample data.
This $\hat{\beta}_1$ has a sampling distribution.

If our four SLR assumptions hold, the sampling distribution of $\hat{\beta}_1$ is normal with mean β_1 and standard deviation which we estimate by

Under H_0: $\beta_1 = 0$, the statistic $\dfrac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$
has a t-distribution with $n - 2$ d.f.

Test for Model Usefulness

One-Tailed Tests		**Two-Tailed Test**
H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$	H_0: $\beta_1 = 0$
H_0: $\beta_1 < 0$	H_0: $\beta_1 > 0$	H_0: $\beta_1 \neq 0$

Test statistic: $t = \dfrac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}}$

Rejection region:

$t < -t_{\alpha}$	$t > t_{\alpha}$	$t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

P-value:

left tail area outside t	right tail area outside t	2*(tail area outside t)

Example: In the drug reaction example, recall $\hat{\beta}_1 = 0.7$.
Is the real β_1 significantly different from 0?
(Use $\alpha = .05$.)

A 100(1 – α)% Confidence Interval for the true slope β_1 is given by:

where $t_{\alpha/2}$ is based on $n - 2$ d.f.

In our example, a 95% CI for β_1 is: