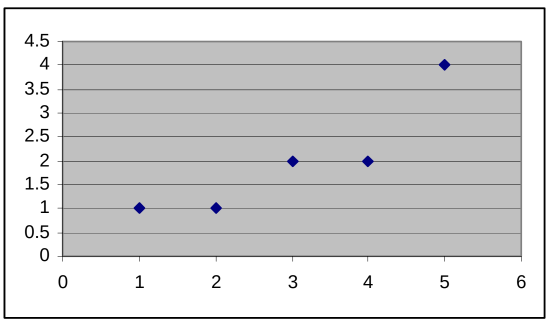
CHAPTER 11 Simple Linear Regression

EXAMPLE

An experiment involving five subjects is conducted to determine the relationship between the percentage of a certain drug in the bloodstream and the length of time it takes the subject to react to a stimulus.

Reaction Time VS. Drug Percentage		
Subject	Amount of Drug Times %	Reaction Time in Seconds
1 Mary	1	1
2 John	2	1
3 Carl	3	2
4 Sara	4	2
5 William	5	4

First recognize that the independent variable **x= Amount of Drug** and the dependent variable **y= Reaction Time .**



The above scatter plot indicate that a model for this situation is the first-order linear model $E(y) = \beta_0 + \beta_1 x$ is probably adequate, we can try to use the sample data to estimate the missing parameters β_1 & β_0 of the least square line..

Preliminary computations for the drug reaction problem					
	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	1	1	1	1	1
	2	1	4	1	2
	3	2	9	4	6
	4	2	16	4	8
	5	4	25	16	20
Totals	$\sum x_i = 15$	$\sum y_i = 10$	$\sum x_i^2 = 55$	$\sum y_i^2 = 26$	$\sum x_i y_i = 37$

The **least squares line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ has the following properties:

1. The sum of errors (SE) equals zero.
2. The sum of squared errors (SSE) is smaller than that for any other straight line model.

Formulas for the Least Squares Estimates
Slope: $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$
y-intercept: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
Where $SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$ **and** $SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$

Example: Find the least squares prediction line for our example above.

Let $E(y) = \beta_0 + \beta_1 x$ be our straight line model where y = reaction time in seconds and x = amount of drug.

Using the numbers from above we can get:

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 37 - (15)(10)/5 = 7$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 55 - \frac{(15)^2}{5} = 55 - 45 = 10$$

then $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 7/10 = 0.7$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\left(\sum y_i\right)}{n} - \hat{\beta}_1 \frac{\left(\sum x_i\right)}{n} = \frac{10}{5} - 0.7\left(\frac{15}{5}\right) = -0.1$

Then the **least squares line** is then given by:
$\hat{y} = -0.1 + 0.7x$

Example: What would the average reaction time be when the % of drug is 2.5%?

Since we found $\hat{y} = -0.1 + 0.7x$, we plug in 2.5 for x to get the predicted average reaction time(y) when the percentage is 2.5%.
is $\hat{y} = -0.1 + 0.7x =$ -0.1 + (0.7)(2.5)=1.65
→ Thus, when the percent of drug is 2.5%, we predict the average reaction time to be 1.65 seconds.

Interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$ of the least square line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
$\hat{\beta}_0$: y intercept, the value of y when x=0 $\hat{\beta}_1$: slope, the mean amount of increase (or decrease) of y for every 1 unit increase in x.

Example: Give practical interpretation to $\hat{\beta}_0$ and $\hat{\beta}_1$ of the example.

We found $\hat{y} = -0.1 + 0.7x$, where $\hat{\beta}_0$= -0.1 and $\hat{\beta}_1$=0.7

- The reaction time (y) is -0.1 seconds when the amount of drug (x) is 0%
- For every 1% increase on the amount of drug (x) in the bloodstream, the mean reaction time (y) is estimated to increase 0.7 seconds.

To find SSE, we have following formula:
$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$
Where $SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$

Example: Now let find the sum squares errors of the least squares prediction line

$$SS_{yy} = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n} = 26 - \frac{(10)^2}{5} = 26 - 20 = 6$$

then $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$=6-(0.7)(7)=1.1

Model Assumptions
Assumption 1: The mean of the probability distribution for (ε) is 0. That is why $E(y) = \beta_0 + \beta_1 x$. Recall the original model was $y = \beta_0 + \beta_1 x + \varepsilon$.
Assumption 2: The variance for (ε) is a constant denoted by σ^2. Now matter what x value we use in the model the distribution of the random error has the same variance.
Assumption 3: The probability distribution of (ε) is normal.
Assumption 4: The values of (ε) associated with any two observed values of y are independent.

An estimator of σ^2

$$S^2 = \frac{SSE}{Degrees\ of\ Freedom\ for\ Error} = \frac{SSE}{n-2}$$

An estimator of σ

$$S = \sqrt{S^2} = \sqrt{\frac{SSE}{n-2}}$$

S is known as the estimated standard error of the regression model.

Interpretation of s, the estimated standard deviation of (ε):

We expect approximately 95% of the observed y values to lie within 2s of their respective least squares predicted y – values, $\hat{y}$.

Example: Find and interpret the estimate of σ

$$s^2 = \frac{SSE}{n-2} = \frac{1.1}{5-2} = 0.367 \quad \text{so} \quad s = \sqrt{s^2} = \sqrt{0.367} = 0.61$$

Interpretation: We expect approximately 95% of the observed y values to lie within 1.22 of their respective least squares line..

Making Inferences about β_1 our slope

Recall β_1 is our slope for the linear model: $y = \beta_0 + \beta_1 x + \varepsilon$.

If the true value of the slope is equal to zero that means $y = \beta_0 + \beta_1 x + \varepsilon$ becomes $y = \beta_0 + 0 \bullet x + \varepsilon = \beta_0 + \varepsilon$, this means that x has no role in predicting y. If that is the case, our model is not useful. For this reason, we will want to test the claim that the slope is equal to zero. We would like to reject that claim because if we are unable to reject it we have a useless model.

To be able to perform a hypothesis test to make an inference about β_1 (the slope), we need to know the sampling distribution of our estimator $\hat{\beta}_1$.

Sampling Distribution of $\hat{\beta}_1$

If we make the four assumptions about ε (see section 11.3), the sampling distribution of the least squares estimator $\hat{\beta}_1$ of the slope will be normal with mean β_1 (the true slope) and standard deviation $\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}}$

We estimate $\sigma_{\hat{\beta}_1}$ by $s_{\hat{\beta}_1} = \dfrac{s}{\sqrt{SS_{xx}}}$ and refer to this quantity as the **estimated standard error of the least squares slope** $\hat{\beta}_1$ (recall $S = \sqrt{S^2} = \sqrt{\dfrac{SSE}{n-2}}$).

A Test of Model Utility: Simple Linear Regression

Hypothesis	Test Statistics	Rejection Region
$H_o; \beta_1 \le 0$ $H_A; \beta_1 > 0$	t= $\dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{\hat{\beta}_1}{s/\sqrt{SSxx}}$ df=n-2	$t > t_a$
$H_o; \beta_1 \ge 0$ $H_A; \beta_1 < 0$		$t < -t_a$
$H_o; \beta_1 = 0$ $H_A; \beta_1 \ne 0$		$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Example: At the 1% significance level, test the claim that there is a positive linear relationship between Amount of Drug (x) and the Reaction Time (y). Use α=0.05

$$H_o; \beta_1 \le 0$$
$$H_A; \beta_1 > 0$$

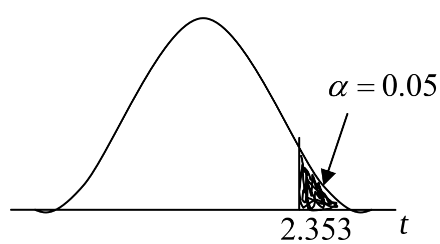
Test Statistics:

$$t = \frac{\hat{\beta}_1}{s/\sqrt{SS_{XX}}} = \frac{0.7}{0.61/\sqrt{10}} = 3.63$$

And df = n-2 = 5-2 = 3

Rejection Region

t > 2.535



Decision: Reject Ho at α=0.05

Conclusion: There is enough evidence to conclude that there is a positive linear relation between Amount of Drug (x) and the Reaction Time (y)

(two additional examples about this section were given in class)

A 100(1-a)% Confidence Interval for the Sample Linear Regression Slope β_1
$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} = \hat{\beta}_1 \pm t_{\alpha/2}\left(\frac{s}{\sqrt{SS_{XX}}}\right)$$ And $t_{\alpha/2}$ is based on (n-2) degrees of freedom.

Example: Find the 95% CI for the slope β_1, the expected change in reaction time(y) for a 1% increase in the amount of drug in the bloodstream (x)

$$\hat{\beta}_1 \pm t_{\alpha/2}\left(\frac{s}{\sqrt{SS_{XX}}}\right) = 0.7 \pm 3.182\left(\frac{0.61}{\sqrt{10}}\right) = 0.7 \pm 0.61$$

$\rightarrow (0.09, 1.31)$

Interpretation: We are 95% confident that the true mean increase in reaction time(y) per additional 1% of drug is between 0.09 and 1.31 seconds.

The Coefficient of Correlation r
The coefficient of correlation, $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$ is a measure of the strength of the linear relationship between two variables x and y. See page 626 to look at scatter plots and the corresponding values for r. People sometimes misinterpret r. Please remember that if r = 0 it does not mean there is no relationship between x and y it just means there does not seem to be a linear relationship between them. Also, if |r| is close to one, it does not means x causes y or that y causes x. It only means there is some linear relationship between the two variables, but the relationship could be due to some other unknown cause.

Example: Find the coefficient of correlation of the Reaction Time(y) VS Amount of drug (x)

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} = \frac{7}{\sqrt{10 \times 6}} = 0.904$$

Since r is positive and near 1 indicates that the reaction tends(y) to increase as the amount of drug(x) in the bloodstream increases, strong positive linear relationship.

The coefficient of determination r^2
Another way to measure the usefulness of the model is to measure the contribution of x in predicting y. To do this, we calculate how much the errors of prediction of y were reduced by using the information provided by x. SS_{YY} = **total sample variation** of the observations around the sample mean for y, and SSE = the **remaining unexplained sample variability** after fitting the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. The **coefficient of determination** is $r^2 = \frac{\text{Explained sample variability}}{\text{Total sample variability}} = \frac{SS_{YY} - SSE}{SS_{YY}} = 1 - \frac{SSE}{SS_{YY}}$ $\rightarrow r^2 = 1 - \frac{SSE}{SS_{YY}}$
Interpretation of r^2
100(r^2)% of the sample variation in y can be explained by using x to predict y in a straight line model.

Example: Find and interpret the coefficient of determination for the drug reaction example.

$$r^2 = 1 - \frac{SSE}{SS_{YY}} = 1 - \frac{1.10}{6} = 0.817 \qquad \text{(using the above formula)}$$

$$r^2 = (0.904)^2 = 0.817 \qquad \text{(using the correlation coefficient r)}$$

Interpretation: 81.7% of the sample variation in reaction time(y) can be explained by using the amount of drug(x) to predict reaction time(y) in a straight line model.

READING FROM MINITAB OUTPUT

Regression Analysis: Time versus Drug

```
The regression equation is
Time = - 0.100 + 0.700 Drug


Predictor      Coef  SE Coef      T      P
Constant    -0.1000   0.6351  -0.16  0.885
Drug         0.7000   0.1915   3.66  0.035


S = 0.605530   R-Sq = 81.7%   R-Sq(adj) = 75.6%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  4.9000  4.9000  13.36  0.035
Residual Error   3  1.1000  0.3667
Total            4  6.0000
```

First, we recognized that the independent variable **x= Amount of Drug** and the dependent variable **y= Reaction Time .**

- To find $\hat{\beta}_0$ and $\hat{\beta}_1$, look under the coefficients column ,

  $\hat{\beta}_0$ =Constant coef= - 0.10 and $\hat{\beta}_1$=Drug coef=0.07.

  OR just look at **`Time = - 0.100 + 0.700 Drug`**, and recognize the $\hat{\beta}_0$ and $\hat{\beta}_1$.

  →So the least square line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ for this example is $\hat{y} = -0.1 + 0.7x$.

- We can also figure out the SSE from the minitab under SS Residual Error, so SSE=1.10.
- The estimator of σ^2, s^2, is the MS of residual error, so s^2=0.3667 .
- The estimator of σ, s, can be found in **`S = 0.605530`**
- The test statistic for making inferences about slope β_1 is under T of Drug and its df=DF Residual Error
  → so we get Test Statistics: t = 3.66 with df=3
- Coefficient of Determination r^2 is clearly give already as 81.7%
- Coefficient of Correlation r, is positive since we know that the slope $\hat{\beta}_1$ is positive, so we just obtain it by $r = (sign)\sqrt{r^2}$, so $r = +\sqrt{0.817} = 0.90388$
- To get the CI for β_1, we use just use the formula $\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$, where β_1=Constant Coef and $s_{\hat{\beta}_1}$ =SE Coef DRUG.