

## Model assumptions

- In fitting a regression model we make four standard assumptions about the random errors  $\epsilon$ :
  1. Mean zero: The mean of the distribution of the errors  $\epsilon$  is zero. If we were to observe an infinite number of  $y$  for each value of  $x$ , then the average of the residuals  $y_i - E(y_i) = 0$ : for a given  $x$ , the mean of  $y$  (or the *expected value of  $y$* ) is equal to  $\beta_0 + \beta_1 x$ .
  2. Constant variance: The variance of the probability distribution of  $\epsilon$  is constant for all  $x$ . We use  $\sigma^2$  to denote the variance of the errors  $\epsilon$ .
  3. Normality: The errors have a normal distribution. Given (1) and (2) above,  $\epsilon \sim N(0, \sigma^2)$ .
  4. Independence: The errors associated to different observations are independent. The deviation of  $y_i$  from the line does not affect the deviation of  $y_j$  from the line.

## Model assumptions (cont'd)

- Three first assumptions are illustrated in Fig. 3.6, page 106 of textbook.
- Why make those assumptions? They make it possible for us to
  1. Develop measures of reliability for regression coefficients
  2. Test hypothesis about the association between  $y$  and  $x$ : draw *inferences*.
- We can test whether the assumptions are met when considering a specific application.
- If assumptions not met, remedies are possible.
- Assumptions often met approximately in real world applications.

## Estimating $\sigma^2$

- The variance of the errors  $\sigma^2$  determines how much observations deviate from their expected value (on the regression line).
- The larger the variance, the larger the deviations (see figure).
- If observations deviate greatly from the line, then the line might not be a good summary of the association between  $y$  and  $x$  after all. Estimators of  $\beta_0$  and  $\beta_1$  and predictions  $\hat{y}$  will be less reliable.
- Typically (essentially always)  $\sigma^2$  is *unknown* and needs to be estimated from the sample, just like the regression coefficients.

## Estimating $\sigma^2$ (cont'd)

- The estimator of  $\sigma^2$  is denoted  $S^2$  or  $\hat{\sigma}^2$  and is obtained as

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

where

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - b_0 - b_1 x_i)^2.$$

- The denominator above is  $n - 2$  because we have used up two degrees of freedom to estimate  $\beta_0$  and  $\beta_1$  out of a total of  $n$  observations.
- $S^2$  is also known as *Mean Squared Error* or MSE.
- The positive square root of  $S^2$  which we denote  $S$  is sometimes known as the *Root Mean Squared Error* or RMSE.

## Estimating $\sigma^2$ (cont'd)

- A different formula that can be used to obtain  $S^2$  is

$$S^2 = \frac{SS_{yy} - b_1 SS_{xy}}{n - 2},$$

where

$$SS_{yy} = \sum_i y_i^2 - n(\bar{y})^2,$$

and  $SS_{xy}$  is the same we described earlier, when obtaining  $b_1$ .

## Estimating $\sigma^2$ - Example

- Consider the advertising expenditures and product sales of five stores. We found  $b_1 = 0.8$ ,  $SS_{xy} = 8$  and  $\bar{y} = 6.8$ .
- We augment data table with  $y_i^2$ :

Store	$x$	$y$	$xy$	$x^2$	$y^2$
1	2	5	10	4	25
2	3	7	21	9	49
3	4	6	24	16	36
4	5	7	35	25	49
5	6	9	54	36	81

so that

$$S_{yy} = (25 + 49 + 36 + 49 + 81) - 5 \times (6.8)^2 = 240 - 5 \times 46.24 = 8.8.$$

## Example (cont'd)

- Then:

$$\begin{aligned} S^2 &= \frac{SS_{yy} - b_1 SS_{xy}}{n - 2} \\ &= \frac{8.8 - 0.8 \times 8}{3} = 0.8 \end{aligned}$$

- Alternatively, we could have used the other formula:

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

(see next page for computation steps)

## Example (cont'd)

1. Compute  $\hat{y}$  for each of the 5 stores:  $\hat{y}_1 = 3.6 + 0.8 \times 2 = 5.2$ ,  
 $\hat{y}_2 = 3.6 + 0.8 \times 3 = 6$ ,  $\hat{y}_3 = 3.6 + 0.8 \times 4 = 6.8$ ,  $\hat{y}_4 = 3.6 + 0.8 \times 5 = 7.6$   
and  $\hat{y}_5 = 3.6 + 0.8 \times 6 = 8.4$ .
2. Compute  $y_i - \hat{y}_i$  for each store:  $5 - 5.2 = -0.2$ ,  $7 - 6 = 1$ ,  $6 - 6.8 = -0.8$ ,  
 $7 - 7.6 = -0.6$  and  $9 - 8.4 = 0.6$ .
3. Get SSE as the sum of the squared deviations:  $SSE = 0.04 + 1 + 0.64 + 0.36 + 0.36 = 2.4$
4. Divide SSE into degrees of freedom  $n - 2$ :  $2.4/3 = 0.8$

## Estimating $\sigma^2$ (cont'd)

- All computer programs produce both SSE and  $S^2$ .
- In Tampa home sales example, there were 92 observations and 90 degrees of freedom for error. From JMP and from SAS output:

$$SSE = 96,746$$

$$MSE = S^2 = 1,074.95$$

$$RMSE = S = 32.79.$$

## Estimating $\sigma^2$ (cont'd)

- Remember that the regression line gives the mean of  $y$  for each value of  $x$ .
- Also, we assume that errors are normally distributed.
- Then most observed  $y$  values will lie within  $\pm 2 \times RMSE$  of the fitted regression line.
- In Tampa home sales example, most sale values will lie within  $\pm 2 \times 32.79 = \pm 66$  (in \$1,000) of the expected sale price for a given assessed value.

## Relative magnitude of errors

- When is the MSE (or  $S^2$ ) too large to make the results from regression useless?
- If  $S^2$  is large relative to the response mean  $\bar{y}$ , then using the results in practice may not be warranted.
- The **coefficient of variation** tells us whether  $S^2$  is small enough or too large:

$$CV = 100(S/\bar{y}).$$

## Relative magnitude of errors (cont'd)

- In application, we hope to see a  $CV$  of 20% or smaller to be assured that the model will produce reliable predictions  $\hat{y}$ .
- In advertising expenditures example,  $CV = 100(0.8944/6.8) = 13.15\%$ .
- In Tampa sales example,  $CV = 100(32.79/236.56) = 13.86\%$ .
- In both examples, the regression model is useful in that it will produce accurate prediction of  $y$  for a given value of  $x$ .
- SAS produces an estimate of the CV.

## Making inferences about the slope $\beta_1$

- If  $x$  does not contribute information about the response  $y$ , then we cannot predict  $y$  given a certain value for  $x$ .
- If sale price of homes in Tampa has no relationship to their assessed value, then assessed value is not a good predictor of sale price when home is put up for sale.
- If so, the expected value of  $y$  is *not* a linear function of  $x$  as postulated by the model:

$$E(y) = \beta_0 + \beta_1 x.$$

- See figure.
- When  $x$  contributes no information about  $y$ , the true slope  $\beta_1$  is equal to zero.

## Inferences about the slope (cont'd)

- We wish to test the hypothesis that in fact  $\beta_1 \neq 0$ .
- To do so, we set up a *test of hypotheses*:
  - We formulate two competing hypothesis
  - We check to see which of the two is supported by the data.
- The two competing hypothesis are called *null* and *alternative*:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0.$$

- If data support  $H_a$ , we conclude that  $x$  does contribute information about  $y$  and therefore that the regression model is useful to make predictions.

## Hypothesis testing for the slope

- To decide whether the data support  $H_a$  we need to know the *sampling distribution* of  $b_1$ .
- The sampling distribution of  $b_1$  is normal, with mean  $\beta_1$  and variance  $\sigma_{b_1}^2$ :

$$b_1 \sim N(\beta_1, \sigma_{b_1}^2)$$

- The variance of  $b_1$  is estimated as

$$\hat{\sigma}_{b_1}^2 = \frac{MSE}{SS_{xx}}$$

## Hypothesis testing for the slope (cont'd)

- In stores example:

$$\hat{\sigma}_{b_1}^2 = \frac{0.8944}{10} = 0.08944.$$

and therefore the *standard error* of  $b_1$  is  $\sqrt{0.08944} = 0.2991$ .

- The standard error for the slope is given by SAS and JMP in the output under label 'Std error' (JMP) or 'Standard error' (SAS) on the row corresponding to  $b_1$ .
- In Tampa sales example, the standard error of  $b_1$  is 0.027.

## A confidence interval for $\beta_1$

- Given a standard error for  $b_1$ , we can compute a confidence interval for the true slope as we did in Chapter 1 (for  $\mu$ ): A 95% confidence interval for the true slope  $\beta_1$  is given by

$$b_1 \pm 2 \times t_{\frac{\alpha}{2}, n-2} \times \sigma_{b_1}.$$

- In Tampa sales example, a 95% CI for the true slope is

$$1.07 \pm 2 \times 0.027 = (1.016, 1.124).$$

## Hypothesis test for $\beta_1$ (cont'd)

- To decide between  $H_0$  and  $H_a$ , we proceed as follows:
  1. Compare the observed  $b_1$  to the hypothesized value  $\beta_1 = 0$ .
  2. If the difference  $b_1 - 0$  is sufficiently small, then the data would suggest that in fact the true value of the slope is zero.
  3. If, however, the difference  $b_1 - 0$  is very large, then the data would suggest that the true value of the slope is *not* zero.

## Hypothesis test for the slope (cont'd)

- How small is small?
- To decide whether  $b_1 - 0$  is sufficiently small to reject  $H_a$ , we re-express the difference in standard error units:

$$\frac{b_1 - 0}{\text{Standard error of } b_1}$$

- But the quantity above looks like a  $z$  (or a  $t$ ) random variable!
- Why????? Remember Chapter 1!

- Since  $b_1 \sim N(\beta_1, \sigma_{b_1}^2)$ , then

$$\frac{b_1 - \beta_1}{\sigma_{b_1}} = z,$$

with  $z \sim N(0, 1)$  as in Chapter 1.

- It appears that we might be able to use the  $z$ -table to make statements about  $\beta_1$  such as: the probability that  $\beta_1 = 0$  is very small or something to that effect.
- Because we do not know the true value of  $\sigma_{b_1}$  we will use the  $t$ -table rather than the  $z$ -table, as we did in Chapter 1.

## Hypothesis test for the slope (cont'd)

- So far we have:
  1. We hypothesize the  $\beta_1 = 0$ .
  2. We check to see whether the data suggest that our null hypothesis holds:

$$t = \frac{b_1 - 0}{\text{standard error of } b_1}$$

3. If  $t$  is large, then the data suggest that  $\beta_1 \neq 0$  and we conclude  $H_a$ .
  4. If  $t$  is small, then we reject  $H_a$  and conclude that there is no evidence in favor of  $H_a$ .
- We need a *critical value* to decide whether our  $t$  is small or large.

## Hypothesis test for the slope (cont'd)

- To get the critical value, we use the  $t$ -table.
- We get the  $t$ -value from the table for a given confidence  $(1 - \alpha)$  and degrees of freedom  $n - 2$ .
- To test the hypothesis that  $\beta_1$  is zero at the 95% confidence level, we make the following decision:

$$\begin{aligned} \text{If } t \leq t_{\frac{\alpha}{2}, n-2} & \quad \text{reject } H_0 \\ \text{If } t > t_{\frac{\alpha}{2}, n-2} & \quad \text{conclude } H_0. \end{aligned}$$

## Tampa sales example

- In Tampa sales example:  $b_1 = 1.068$  and  $\hat{\sigma}_{b_1} = 0.02709$ .
- We wish to test whether  $\beta_1$  is different from zero at the 95% confidence level.
- Calculate  $t = \frac{b_1 - 0}{\hat{\sigma}_{b_1}} = \frac{1.068}{0.02709} = 39.42$ .
- With  $\alpha/2 = 0.025$  and  $n - 2 = 90$  degrees of freedom, the critical value from the  $t$ -table is 1.98 (using the row corresponding to 120 df).
- Since  $39.42 > 1.98$  we conclude  $H_a$ : there is evidence to conclude that the true slope  $\beta_1$  is different from zero.

## Advertising expenditures example

- In advertising expenditures example:  $b_1 = 0.8$ ,  $\hat{\sigma}_{b_1}^2 = 0.08944$ .
- We wish to test whether  $\beta_1$  is different from zero at the 95% confidence level.
- Calculate:  $t = \frac{b_1 - 0}{\hat{\sigma}_{b_1}} = \frac{0.8}{\sqrt{0.08944}} = 2.675$ .
- From the  $t$ -table, with  $\alpha/2 = 0.025$  and  $n - 2 = 3$  we get: 3.182.
- Since  $2.675 < 3.182$  we reject  $H_a$ : we have no evidence to conclude that the true  $\beta_1 \neq 0$ .
- With a confidence level of 90%, the table value is 2.353 and we conclude  $H_a$ . However, we have a 10% chance of having reached the wrong conclusion.