Coefficient of Determination

- The coefficient of determination R^2 (or sometimes r^2) is another measure of how well the least squares equation

$$\hat{y} = b_0 + b_1 x$$

performs as a predictor of y.

- R^2 is computed as:

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SS_{yy}}{SS_{yy}} - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

- R^2 measures the relative sizes of SS_{yy} and SSE. The smaller SSE, the more reliable the predictions obtained from the model.

Coefficient of Determination (cont'd)

- The higher the R^2, the more useful the model.
- R^2 takes on values between 0 and 1.
- Essentially, R^2 tells us how much better we can do in predicting y by using the model and computing $\hat{y}$ than by just using the mean $\bar{y}$ as a predictor.
- Note that when we use the model and compute $\hat{y}$ the prediction depends on x because $\hat{y} = b_0 + b_1 x$. Thus, we act as if x contains information about y.
- If we just use $\bar{y}$ to predict y, then we are saying that x does not contribute information about y and thus our predictions of y do not depend on x.

Coefficient of Determination (cont'd)

- More formally:
  - SS_{yy} measures the deviations of the observations from their mean: $SS_{yy} = \sum_i (y_i - \bar{y})^2$. If we were to use $\bar{y}$ to predict y, then SS_{yy} would measure the variability of the y around their predicted value.
  - SSE measures the deviations of observations from their predicted values: $SSE = \sum_i (y_i - \hat{y}_i)^2$.

- If x contributes no information about y, then SS_{yy} and SSE will be almost identical, because $b_1 \approx 0$.

- If x contributes lots of information about y then SSE is very small.

- Interpretation: R^2 tells us how much better we do by using the regression equation rather than just $\bar{y}$ to predict y.

Coefficient of Determination - Example

- Consider Tampa sales example. From printout, $R^2 = 0.9453$.
- Interpretation: 94% of the variability observed in sale prices can be explained by assessed values of homes. Thus, the assessed value of the home contributes a lot of information about the home's sale price.
- We can also find the pieces we need to compute R^2 by hand in either JMP or SAS outputs:
  - SS_{yy} is called Sum of Squares of Model in SAS and JMP
  - SSE is called Sum of Squares of Error in both SAS and JMP.
- In Tampa sales example, $SS_{yy} = 1673142$, $SSE = 96746$ and thus

$$R^2 = \frac{1673142 - 96746}{1673142} = 0.94.$$

Estimation and prediction

- With our regression model, we might wish to do two things:

  1. Estimate the mean (or expected) value of y for a given x.
  2. Predict the value of a single y given a value of x.

- In both cases, we use the same sample estimator (or predictor):

$$\hat{y} = b_0 + b_1 x.$$

- The difference between estimating a mean or predicting a single observation is in the accuracy with which we can do each of these two things – the standard errors in each of the two cases are different.

Estimation and prediction (cont'd)

- The standard deviation of the estimator $\hat{y}$ of the **mean** of y for a certain value of x, say x_p is

$$\sigma_{\hat{y}} = \sigma \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}},$$

  where

  - σ is the error standard deviation, estimated by RMSE (or S).
  - x_p is the specific value of x for which we wish to estimate the mean of the y

- $\sigma_{\hat{y}}$ is called the **standard error of** $\hat{y}$.

- If we use RMSE in place of σ, we obtain an estimate $\hat{\sigma}_{\hat{y}}$.

Estimation and prediction (cont'd)

- The standard deviation of the estimator $\hat{y}$ of an **individual** $y-$value given a certain value of x, say x_p is

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}},$$

- We call $\sigma_{(y-\hat{y})}$ the **standard error of prediction**.

- If we use $RMSE$ (or S) in place of σ, then we have an estimate of the standard error of prediction, and we denote the estimate by $\hat{\sigma}_{(y-\hat{y})}$.

Estimation and prediction - Example

- Consider the Tampa sales example, and refer to the JMP ouput. From output, $RMSE == S = 32.78$, and mean assessed price ($\bar{x}$) is \$201.75.

- We wish to estimate the mean price of houses assessed at $x_p = \$320$ (in \$1,000s) and also compute $\hat{\sigma}_{\hat{y}}$, the standard error of $\hat{y}$:

$$\hat{y} = 20.94 + 1.069 \times 320 = 363.$$

- To compute $\hat{\sigma}_{\hat{y}}$ we also need SS_{xx}. We use the computational formula

$$SS_{xx} = \sum_i x_i^2 - n(\bar{x})^2.$$

Estimation and prediction - Example

- To get $\sum_i x_i^2$ we can create a new column in JMP which is equal to Assvalue squared, and then ask for its sum.

- In Tampa sales example:

$$SS_{xx} = \sum_i x_i^2 - n(\bar{x})^2 = 5,209,570.75 - 92 \times 201.75^2 = 1,464,889.$$

- An estimate of the standard error of $\hat{y}$ is now:

$$\begin{aligned} \hat{\sigma}_{\hat{y}} &= 32.78\sqrt{\frac{1}{92} + \frac{(320 - 201.75)^2}{1,464,889}} \\ &= 32.78\sqrt{0.01086 + 0.009545} \\ &= 4.68. \end{aligned}$$

Estimation and prediction - Example

- Suppose that now we wish to predict the sale price of a single house that is appraised at \$320,000.

- The point estimate is the same as before: $\hat{y} = 20.94 + 1.069 \times 320 = 363$.

- The **standard error of prediction** however is computed using the second formula:

$$\hat{\sigma}_{(y-\hat{y})} = S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}.$$

Estimation and prediction - Example

- We have S (or $RMSE$), n, $(x_p - \bar{x})^2$ and SS_{xx} from before, so all we need to do is

$$\begin{aligned}
\hat{\sigma}_{(y-\hat{y})} &= 32.78\sqrt{1 + \frac{1}{92} + \frac{(320 - 201.75)^2}{1,464,889}} \\
&= 32.78\sqrt{1 + 0.01086 + 0.009545} \\
&= 33.11
\end{aligned}$$

Estimation and prediction - Example

- Note that in Tampa sales example, $\hat{\sigma}_{(y-\hat{y})} > \hat{\sigma}_{\hat{y}}$ (33.11 versus 4.68).

- This is true always: we can estimate a mean value for y for a given x_p much more accurately than we can predict the value of a single y for $x = x_p$.

  - In estimating a mean y for $x = x_p$, the only uncertainty arises because we do not know the $true$ regression line.
  - In predicting a single y for $x = x_p$, we have two uncertainties: the $true$ regression line plus the expected variability of $y-$values around the true line.

Estimation and prediction - Using JMP

- For each observation in a dataset we can get from JMP (or from SAS): $\hat{y}$, $\hat{\sigma}_{\hat{y}}$, and also $\hat{\sigma}_{(y-\hat{y})}$.

- In JMP do:

  1. Choose *Fit Model*
  2. From *Response* icon, choose *Save Columns* and then choose *Predicted Values*, *Std Error of Predicted*, and *Std Error of Individual*.

Estimation and prediction - Using JMP

- A **VERY unfortunate thing!** JMP calls things different from the book:
  - In book: $\hat{\sigma}_{\hat{y}}$ is standard error of estimation but in JMP it is standard error of prediction.
  - In book: $\hat{\sigma}_{(y-\hat{y})}$ is standard error of prediction but in JMP it is standard error of individual.

- SAS calls them the same as the book: standard error of the mean and standard error of prediction.

Confidence intervals

- We can compute a $100(1-\alpha)\%$ CI for the **true mean** of y at $x = x_p$.

- We can also compute a $100(1-\alpha)\%$ CI for **true value of a single** y at $x = x_p$.

- In both cases, the formula is the same as the general formula for a CI:

$$\text{estimator} \pm t_{\frac{\alpha}{2}, n-2} \text{ standard error}$$

Confidence intervals (cont'd)

- The CI for the **true mean** of y at $x = x_p$ is

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{\hat{y}}$$

- The CI for a **true single** value of y at $x = x_p$ is

$$\hat{y} \pm t_{\frac{\alpha}{2},n-2}\hat{\sigma}_{(y-\hat{y})}$$

Confidence intervals - Example

- In Tampa sales example, we computed $\hat{y}$ for $x = 320$ and we also computed the standard error of the mean of y and the standard error of a single y at $x = 320$.

- The 95% CI for the true mean of y at $x = 320$ is

$$\begin{aligned} 95\% CI &= \hat{y} \pm t_{\frac{\alpha}{2}, n-2} \hat{\sigma}_{\hat{y}} \\ &= 363 \pm 1.98 \times 4.68 = (354,\ 372). \end{aligned}$$

- The 95% CI for the true value of a single y at $x = 320$ is

$$\begin{aligned} 95\% CI &= \hat{y} \pm t_{\frac{\alpha}{2}, n-2} \hat{\sigma}_{(y-\hat{y})} \\ &= 363 \pm 1.98 \times 33.11 = (297,\ 429). \end{aligned}$$

Confidence intervals - Interpretation

- The 95% CI for the mean sale price of houses assessed at \$320,000 is \$354,000 to \$372,000. If many houses assessed at about \$320,000 go on the market, we expect that the mean sale price of those houses will be included within those two values.

- The 95% CI for the sale price of a single house that is assessed at \$320,000 is \$297,000 to \$429,000. That means that a homeowner who has a house valued at \$320,000 can expect to get between \$297,000 and \$429,000 if she decides to sell the house.

- Again, notice that it is much more difficult to precisely predict a single value than it is to predict the mean of many values.

- See Figure 3.25 on page 135 of textbook.