An Estimator of σ

Recall the assumption for least squares regression: ε are normally distributed with mean 0 and standard deviation σ

Reasonable to assume that the greater the variability in the random error:

- Greater the errors in the estimation of the model parameters β_0 and β_1
- Greater the error of prediction when $\hat{y}$ is used to predict y for some value of x

σ is usually unknown and must be estimated from the data

The best estimate is denoted s_e and is obtained by:

$$s_e = \sqrt{\frac{\text{SSE}}{\text{degrees of freedom for error}}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Recall:

$$\text{SSE} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left[y_i - \left(b_0 + b_1 x_i\right)\right]^2$$

s_e is termed the ***estimated standard error of the regression model***

Recalling the drug reaction example:

x	y	$\hat{y} = -.1 + .7x$	$(y-\hat{y})$	$(y-\hat{y})^2$
1	1	.6	(1-.6) = .4	.16
2	1	1.3	(1-1.3) = -.3	.09
3	2	2.0	(2-2.0) = 0	.00
4	2	2.7	(2-2.7) = -.7	.49
5	4	3.4	(4-3.4) = .6	.36
			Sum of errors = 0	Sum of squared errors (SSE) = 1.10

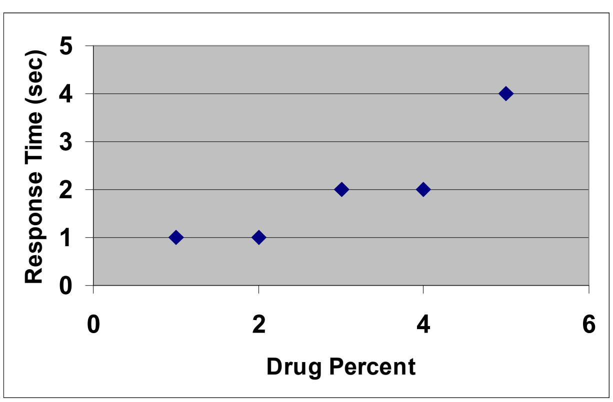
$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{1.10}{3}} = \sqrt{0.367} = 0.61$$

Interpretation

We expect most (approximately 95%) of the observed y values to lie within $2s_e$ of $\hat{y}$, their respective least squares predicted values

Assessing the Utility of the Model: Making Inferences about the Slope

Suppose that the reaction times, y, are completely unrelated to the percentage of drug in the blood stream.



What could be said about the values of β_0 and β_1 in the probabilistic model $y = \beta_0 + \beta_1 x + \varepsilon$ if x contributes no information for the prediction of y?

This means that the deterministic part of the model

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

does not change as x changes

For the linear model, this means that the true slope, β_1, must be equal to zero.

We can test set up a hypothesis test for this:

H_0: $\beta_1 = 0$; the linear model contains no information for the prediction of y

H_1: $\beta_1 \neq 0$; the linear model is useful for predicting y

Choice of test statistic is found by considering the sampling distribution of b_1, the least squares estimator of the slope, β_1

Sampling Distribution of b_1

Given ε distributed iid $N(0,\sigma)$, the sampling distribution of b_1 will be normal with:

$$E(b_1) = \beta_1$$

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\mathrm{SS}_{xx}}}$$

σ_{b_1} is estimated by $s_{b_1} = \dfrac{s_e}{\sqrt{\mathrm{SS}_{xx}}}$

s_{b_1} is call the estimated standard error of the least squares slope b_1

The test statistic is the *t*-statistic

$$t = \frac{b_1 - \text{ hypothesized value of } \beta_1}{s_{b_1}}$$

where

$$s_{b_1} = \frac{s_e}{\sqrt{\mathrm{SS}_{xx}}}$$

By substitution,

$$t = \frac{b_1 - 0}{s_e / \sqrt{\mathrm{SS}_{xx}}}$$

Test of Model Utility: Simple Linear Regression

One-tailed test	**Two-tailed test**
$H_0 : \beta_1 \geq 0$ $H_1 : \beta_1 < 0$	$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$
$t = \dfrac{b_1 - 0}{s_e / \sqrt{\mathrm{SS}_{xx}}}$	$t = \dfrac{b_1 - 0}{s_e / \sqrt{\mathrm{SS}_{xx}}}$
$t < -t_{\alpha(1)}$ (or $t > t_{\alpha(1)}$ when $H_1 : \beta_1 > 0$)	$t < -t_{\alpha(2)}$ or $t > t_{\alpha(2)}$

$t_{\alpha(1)}$ and $t_{\alpha(2)}$ are based on $(n-2)$ df

<u>Assumptions</u>

mean of $\varepsilon = 0$;
standard deviation of $\varepsilon = \sigma$ (constant);
ε are normally distributed;
values of ε are independent

For the drug response example define the hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Let $\alpha = .05$, $n = 5 \Rightarrow 5 - 2 = 3$ df

Rejection region: $|t| > t_{\alpha(2),\mathrm{df}} = t_{.05(2),3} = 3.182$

$b_1 = .7$; $s_e = .61$; $\mathrm{SS}_{xx} = 10$

$$t = \frac{b_1 - 0}{s_e / \sqrt{\mathrm{SS}_{xx}}} = \frac{.7}{.61/\sqrt{10}} = \frac{.7}{.19} = 3.7$$

Conclusion?

Reject H_0 and conclude that there is sufficient evidence to support the alternative hypothesis that the slope β_1 is not 0.

The sample evidence indicates that the amount of drug, x, in the blood stream contributes information for the prediction of reaction time, y, when a linear model is used

What conclusion would we have drawn if t did not fall in the rejection region (P-value $> \alpha$)?

We fail to reject H_0---do not conclude that $\beta_1 = 0$.

- Additional data may indicate that $\beta_1 \neq 0$
- The relationship between x and y may be more complex and require fitting of another model

$100(1-\alpha)\%$ Confidence Interval for the Simple Linear Regression Slope β_1

$$b_1 \pm \underbrace{t_{\alpha(2),df}\, s_{b_1}}_{E}$$

where the estimated standard error of b_1 is calculated by

$$s_{b_1} = \frac{s_e}{\sqrt{\mathrm{SS}_{xx}}}$$

$t_{\alpha(2)}$ is based on $n-2$ degrees of freedom

Same assumptions as for hypothesis tests for b_1

For the drug response example the 95% confidence interval for the slope, β_1

$$\begin{aligned} b_1 \pm t_{\alpha(2),df}\, s_{b_1} = b_1 \pm t_{.05(2),3}\, s_{b_1} &= .7 \pm 3.182\left(\frac{s_e}{\sqrt{\mathrm{SS}_{xx}}}\right) \\ &= .7 \pm 3.182\left(\frac{.61}{\sqrt{10}}\right) \\ &= .7 \pm .61 \\ &\Rightarrow (.09, 1.31) \end{aligned}$$

Interpretation?

We can be 95% confident that the true mean increase in reaction time per additional 1% of the drug is between .09 and 1.31 seconds.

This inference is only meaningful over the sampled drug range of 1% to 5%. Since all the values in this interval are positive, it appears that β_1 is positive and that the mean of y increase as x increases.

1. Select Tools (or Data) > Data Analysis…
2. Select Regression
3. Input Y range and X range—if you included Column labels in your data ranges, check Labels; Check Confidence Level (default is 95%); might also want to check Residual Plots and Normal Probability Plots

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.904
R Square	**0.817**
Adjusted R Square	0.756
Standard Error	**0.606**
Observations	**5**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4.9	4.900	13.364	0.0354
Residual	3	1.1	0.367		
Total	4	6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**-0.1**	**0.635**	**-0.157**	**0.8849**	**-2.121**	**1.921**
DrugPct	**0.7**	**0.191**	**3.656**	**0.0354**	**0.091**	**1.309**

Using the LS Model for Estimation and Prediction

Estimation ➔ use of model for estimating the mean value of y, $E(y)$, for a specific value of x.

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

Fitted line gives us $\hat{y} = b_0 + b_1 x$ and this implies $(\hat{y}|x = x_i) = b_0 + b_1 x_i$

Estimation gives the mean response for *all* values of y for a given value of x

In the drug reaction example, the regression equation gives us an estimate of the mean response time for all people whose blood contain x% of the drug

Prediction ➔ use of the model for predicting a new individual y value for a given x value

In the drug reaction example, one may want to predict the reaction time for a specific person who possesses 4% of the drug in the bloodstream

In estimation, we are attempting to estimate the mean value of y for a very large number of experiments at the given x value

In prediction, we are trying to predict the outcome, y, for a single experiment at the given x value

We use the LS prediction equation, $\hat{y} = b_0 + b_1 x$, for both estimation of the mean value of y and for prediction of a specific new value of y

We can create ***confidence*** and ***prediction intervals*** for estimations and predictions.

Standard error of $\hat{y}$, $\sigma_{\hat{y}}$

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_{xx}}}$$

Standard error of prediction, $\sigma_{(y-\hat{y})}$

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_{xx}}}$$

where σ is the standard deviation of the random error ε--use s_e as an estimate of σ; you will need $\bar{x}$ and SS_{xx}

A $100(1-\alpha)$% Confidence Interval for the mean value of y at $x = x_0$

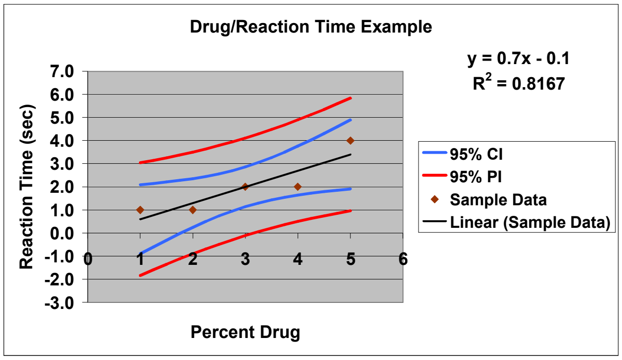
$$\hat{y} \pm t_{\alpha(2),df}\sigma_{\hat{y}} = \hat{y} \pm t_{\alpha(2),(n-2)} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_{xx}}}$$

(formula in Triola, problem 26 page 566)

A $100(1-\alpha)$% Prediction Interval for an individual new value of y at $x = x_0$

$$\hat{y} \pm t_{\alpha(2),df}\sigma_{\hat{y}} = \hat{y} \pm t_{\alpha(2),(n-2)} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\mathrm{SS}_{xx}}}$$

(formula in Triola, page 561)



DrugPct	RepSec	x-xbar	(x-xbar)^2
1	1	-2	4
2	1	-1	1
3	2	0	0
4	2	1	1
5	4	2	4
3			10

tcrit	3.182
s	0.606

xnot	yhat	E-CI	E-PI	CILow	CIUp	PILow	PIUp
1	0.6	1.494	2.439	-0.894	2.094	-1.839	3.039
1.25	0.775	1.372	2.367	-0.597	2.147	-1.592	3.142
1.5	0.95	1.257	2.302	-0.307	2.207	-1.352	3.252
1.75	1.125	1.151	2.246	-0.026	2.276	-1.121	3.371
2	1.3	1.056	2.199	0.244	2.356	-0.899	3.499
2.25	1.475	0.976	2.161	0.499	2.451	-0.686	3.636
2.5	1.65	0.915	2.134	0.735	2.565	-0.484	3.784
2.75	1.825	0.876	2.118	0.949	2.701	-0.293	3.943
3	2	0.862	2.112	1.138	2.862	-0.112	4.112
3.25	2.175	0.876	2.118	1.299	3.051	0.057	4.293
3.5	2.35	0.915	2.134	1.435	3.265	0.216	4.484
3.75	2.525	0.976	2.161	1.549	3.501	0.364	4.686
4	2.7	1.056	2.199	1.644	3.756	0.501	4.899
4.25	2.875	1.151	2.246	1.724	4.026	0.629	5.121
4.5	3.05	1.257	2.302	1.793	4.307	0.748	5.352
4.75	3.225	1.372	2.367	1.853	4.597	0.858	5.592
5	3.4	1.494	2.439	1.906	4.894	0.961	5.839

The prediction interval for an individual new value of *y* is always wider than the corresponding CI for the mean value of *y*

Error of estimation and error of prediction take their smallest values when $x_0 = \bar{x}$.

The farther x_0 lies from $\bar{x}$, the larger will be the errors of estimation and prediction---this is because the deviation is larger at the extremes of the interval where the largest and smallest values of *x* in the data set occur

The CI width grows smaller as *n* is increased---in theory you can obtain as precise an estimate of the mean value of *y* (for a given value of *x*) as desired by selecting a large enough sample

The prediction interval for a new value of *y* also gets smaller as *n* increases, but there is a lower limit on its width---the interval can get no smaller than $\hat{y} \pm z_{\alpha(2)} s_e$

The only way to obtain more accurate predictions for new values of *y* is to reduce the standard deviation of the regression model---improve the model

Problem 11.77, page 568 (McClave and Sincich, 9th Edition)

A study was conducted to determine whether a student's final grade in an introductory sociology course is linearly related to his or her performance on the verbal ability tests administered before college entrance.

a) Find the least squares line relating *y* to *x*

b) Plot the data points and graph the least squares line

c) Do the data provide sufficient evidence to indicate that a positive correlation exists between verbal score and final grade?

d) Find a 95% CI for the slope, β_1

e) Predict a student's final grade in the introductory course when his or her verbal test score is 50. Use a 95% prediction interval.

f) Find a 95% CI for the mean final grade for all students scoring 50 on the college entrance verbal exam.

Student	Verbal Ability Test Score (x)	Final Sociology Grade (y)
1	39	65
2	43	78
3	21	52
4	64	82
5	57	92
6	47	89
7	28	73
8	75	98
9	34	56
10	52	75

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.840
R Square	**0.705**
Adjusted R Square	0.668
Standard Error	**8.704**
Observations	**10**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1449.974	1449.974	19.141	0.002
Residual	8	606.026	75.753		
Total	9	2056.000			

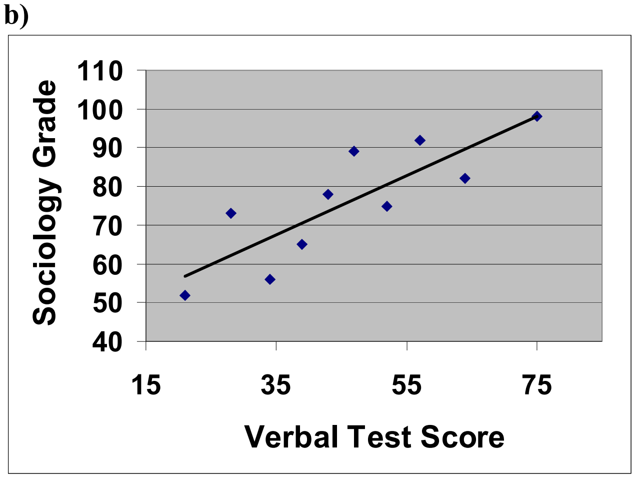
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**40.78**	**8.51**	**4.79**	**0.00**	**21.17**	**60.40**
Verbal	**0.77**	**0.17**	**4.38**	**0.00**	**0.36**	**1.17**

RESIDUAL OUTPUT

Observation	*Grade*	*Residuals*
1	70.64	-5.64
2	73.70	4.30

a) $\hat{y} = 40.78 + 0.77x$

b)



c) Test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 > 0$

$$t = \frac{0.77 - 0}{0.17} = 4.53; \quad t_{.05(1),8} = 1.860$$ ➔ **reject the null**

Conclude that there is sufficient sample evidence to indicate that there is a positive association between verbal score and final sociology grade (*p* < 0.005).

d) 95% CI for

$$\beta_1: \; b_1 \pm \underbrace{t_{.05(2),8} SE_{b_1}}_{E} \Rightarrow 0.77 \pm 2.306(0.17) \Rightarrow (.38, 1.16)$$

e) 95% PI for individual given $x_0 = 50$

$$\hat{y} \pm t_{\alpha(2),df}\sigma_{\hat{y}} = \hat{y} \pm t_{\alpha(2),(n-2)} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

$\hat{y} = 40.78 + (.77 \times 50) = 79.28$; $t_{.05(2),8} = 2.306$; $s_e = 8.70$; $SS_{xx} = 2474$

Score (x)	x-xbar	(x-xbar)^2
39	-7	49
43	-3	9
21	-25	625
.	.	.
.	.	.
.	.	.
		2474
	xbar= 46	

$$79.28 \pm 2.306(8.7)\sqrt{1 + \frac{1}{10} + \frac{(50 - \bar{x})^2}{SS_{xx}}}$$

$$79.28 \pm 20.0622\sqrt{1 + \frac{1}{10} + \frac{(50 - 46)^2}{2474}}$$

$$79.28 \pm 20.0622(1.052) \Rightarrow 79.28 \pm 21.10$$

$$\Rightarrow (58.18, 100.38)$$

The probability is 0.95 that a randomly selected student from the population will have a final sociology grade between 58.18 and 100.38 if his/her verbal score was 50. –the 95% PI estimates the central 95% of the values of *y* for members of the population with a specified value of *x*.

f) 95% CI for mean given $x_0 = 50$

$$\hat{y} \pm t_{\alpha(2),df}\sigma_{\hat{y}} = \hat{y} \pm t_{\alpha(2),(n-2)} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

$$79.28 \pm 20.0622\sqrt{\frac{1}{10} + \frac{(50-46)^2}{2474}}$$

$$79.28 \pm 20.0622(0.326) \Rightarrow 79.28 \pm 6.55 \Rightarrow (72.73, 85.83)$$

We are 95% confident that the mean final grade in the introductory sociology course will be between 72.73 and 85.83 for all students whose verbal score was 50.

A PI differs conceptually from a CI.

-- A CI estimates an unknown population parameter, which is a numerical characteristic or summary of the population.

-- A PI does not estimate a population parameter; rather, a PI estimates the potential data value for an individual—it describes an interval into which a specific percentage of the population may fall.

<u>Residual Plots</u>—examine for presence or absence of patterns