Section 11.1 - Inference for Multiple Regression

Statistics 104

Autumn 2004

Multiple Regression

Multiple Regression - Regression with 2 or more predictor variables

Example: Roofing Shingle Sales in $n = 26$ sales districts

$y =$ Annual Sales (in 1000 squares)

$x_1 =$ Promotional expenditures (in \$1000)

$x_2 =$ Number of active accounts

$x_3 =$ Number of competing brands

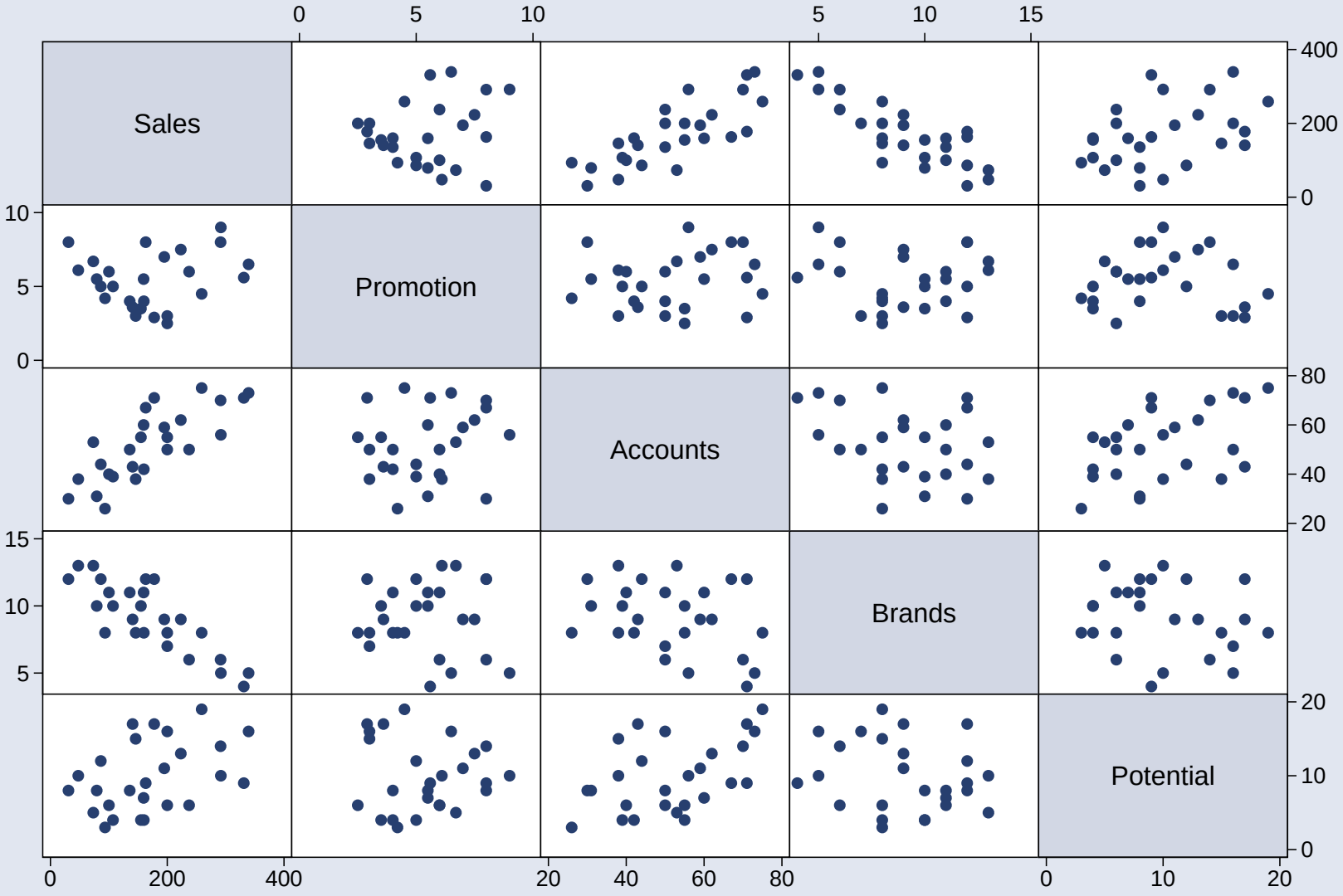
$x_4 =$ District Potential

2 Questions of interest

1. Which of the four factors (if any) affect shingle sales?

2. Does promotion affect sales, after accounting for the other three factors?

Roofing Shingle Sales Data



```
. correlate (obs=26)

             |    sales promot~n accounts   brands potent~l
-------------+---------------------------------------------
       sales |   1.0000
   promotion |   0.1589   1.0000
    accounts |   0.7828   0.1726   1.0000
      brands |  -0.8330  -0.0383  -0.3243   1.0000
   potential |   0.4073  -0.0706   0.4682  -0.2021   1.0000
```

Suggestions from plot and correlation matrix

- Sales increase with number of accounts
- Sales decrease with number of competing brands
- Sales increase with potential
- Correlation among predictors (accounts and promotion, accounts and potiential, etc)

Multiple Linear Regression Model for the Mean Response

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

In the example $p = 4$.

Each combination of predictor variables has its own mean level.

Data

Obs 1: $(x_{11}, x_{12}, \ldots, x_{1p}, y_1)$

Obs 2: $(x_{21}, x_{22}, \ldots, x_{2p}, y_2)$

•

•

•

Obs n: $(x_{n1}, x_{n2}, \ldots, x_{np}, y_n)$

Multiple Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where

$$\epsilon_i \sim N(0, \sigma)$$

Note that no assumption about the distribution of the predictor variables is begin made here.

The model for the y's is conditional on the x's.

How to think about the β_j's

β_j gives the expected change in the response y when x_j increases by 1 unit, given the other predictor variables are help fixed.

For example, suppose that the population mean model for the shingle data is

$$\mu_y = 180 + 2.1\text{Proportion} + 3.3\text{Accounts} - 21\text{Brands} + 0.35\text{Potential}$$

each additional account is worth 3.3 ($\times$ 1000) additional squares of sales on average, assuming that promotional spending, the number of competing brands and sales potential is kept the same.

Another way of thinking of β_j is suppose that you have 2 observation with all the x's the same, except that x_j for observation 1 is 1 unit higher than x_j for observation 2. Then the expected difference in $y_1 - y_2$ is β_j.

Estimation:

- Parameters to be estimated: $\beta_0, \beta_1, \ldots, \beta_p, \sigma$

- Estimation method: Least squares

  Based of the residuals

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip} \end{aligned}$$

  Find $b_0, b_1, \ldots, b_p$ which minimize

$$\sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip})^2$$

  There are not nice formulas for $b_0, b_1, \ldots, b_p$, unless you do things using matrix algebra. (If you are interested, see a regression text like Neter, Kutner, Nachtsheim, & Wasserman or Montgomery and Peck).

- Estimating σ

$$\begin{aligned} s^2 &= \frac{\sum e_i^2}{n-p-1} \\ &= \frac{\sum (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_p x_{ip})^2}{n-p-1} \\ s &= \sqrt{s^2} \end{aligned}$$

Degrees of freedom $= n - p - 1$

Note that the degrees of freedom match with the simple linear regression case ($p = 1$), which gives $n - 2$.

```
. regress sales promotion accounts brands potential
      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  4,    21) =  479.10
       Model |  176777.061     4  44194.2653            Prob > F      =  0.0000
    Residual |  1937.13655    21  92.2445975            R-squared     =  0.9892
-------------+------------------------------           Adj R-squared =  0.9871
       Total |  178714.198    25   7148.5679           Root MSE      =  9.6044

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   promotion |    1.807064   1.081039     1.67   0.109    -.4410806    4.055208
    accounts |    3.317833   .1628917    20.37   0.000     2.979082    3.656585
      brands |   -21.18498   .7879389   -26.89   0.000    -22.82359   -19.54638
   potential |    .3245121   .4677644     0.69   0.495    -.6482572    1.297281
       _cons |    178.3203   12.96032    13.76   0.000     151.3679    205.2728
------------------------------------------------------------------------------
```

$$\hat{\mu}_y = 178.32 + 1.81\text{Proportion} + 3.32\text{Accounts}$$
$$-21.18\text{Brands} + 0.32\text{Potential}$$

Inference on Individual βs

The confidence intervals for the individual βs are similar to the simple linear regression case.

The confidence interval for β_j is

$$b_j \pm t^* SE_{b_j}$$

where t^* is based on $n - p - 1$ degrees of freedom.

Note that the degrees of freedom is given in the Residual (or error) line of the ANOVA table.

For example, a 95% CI for the effect of each additional account in a region is given by

$$\begin{aligned} CI &= 3.318 \pm 2.080 \times 0.1629 \\ &= 3.318 \pm 0.339 \\ &= (2.979, 3.657) \end{aligned}$$

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
promotion	1.807064	1.081039	1.67	0.109	-.4410806	4.055208
accounts	3.317833	.1628917	20.37	0.000	2.979082	3.656585
brands	-21.18498	.7879389	-26.89	0.000	-22.82359	-19.54638
potential	.3245121	.4677644	0.69	0.495	-.6482572	1.297281
_cons	178.3203	12.96032	13.76	0.000	151.3679	205.2728

Tests examining $H_0 : \beta_j = 0$

Interested in only a single β. Assume that the others can take any arbitrary value.

Want to compare two models:

- Full model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{i,p-1} + \beta_p x_{ip} + \epsilon_i$$

- Reduced model (for example $H_0 : \beta_p = 0$):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{i,p-1} + \epsilon_i$$

  (Similarly for $H_0 : \beta_j = 0$)

Investigates the question, is the full model a better description of the data than the reduced model.

Another way of thinking of this setup, does variable j add anything significant to the prediction after using the other $p-1$ variables.

In the simple regression case, the two models compared are

- Full model ($H_A : \beta_1 \neq 0$):

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

- Reduced model ($H_0 : \beta_1 = 0$):

$$y_i = \beta_0 + \epsilon_i$$

To test $H_0 : \beta_j = 0$, a t test can be used

Test statistic:

$$t = \frac{b_j}{SE_{b_j}}$$

P-values:

$$H_A : \beta_1 < 0 \qquad p\text{−value} = P[T \leq t_{obs}]$$
$$H_A : \beta_1 > 0 \qquad p\text{−value} = P[T \geq t_{obs}]$$
$$H_A : \beta_1 \neq 0 \qquad p\text{−value} = 2 \times P[T \geq |t_{obs}|]$$

where T has a $t(n-p-1)$ distribution.

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
promotion	1.807064	1.081039	1.67	0.109	-.4410806	4.055208
accounts	3.317833	.1628917	20.37	0.000	2.979082	3.656585
brands	-21.18498	.7879389	-26.89	0.000	-22.82359	-19.54638
potential	.3245121	.4677644	0.69	0.495	-.6482572	1.297281
_cons	178.3203	12.96032	13.76	0.000	151.3679	205.2728

For promotion it appears, that after accounting for the effects of the number of accounts, the number of brands, and the potential for the region, the amount spent on promotion doesn't significantly affect the sale of shingles. Though note that the sample size is small and the CI for β_{promo} is wide.

Tests investigating all βs

A second set of hypotheses of interest is

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \quad \text{vs} \quad H_A : \text{at least one of the } \beta_j \neq 0$$

The null hypothesis states that none of the predictor variables is useful in describing the response variable.

The alternative hypothesis states that at least one of the predictors is useful (but doesn't specify which of them are).

In the framework of comparing two models, these hypotheses correspond to

- Full model (H_A):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

- Reduced model ($H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$):

$$y_i = \beta_0 + \epsilon_i$$

These two hypotheses can be examined with an ANOVA style analysis

ANOVA Table

Source	DF	SS	MS	F
Model	p	$SSM = \sum(\hat{y}_i - \bar{y})^2$	$MSM = \frac{SSM}{DFM}$	$F = \frac{MSM}{MSE}$
Error	$n - p - 1$	$SSE = \sum(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{DFE}$	
Total	$n - 1$	$SST = \sum(y_i - \bar{y})^2$		

The F statistic should be compared to an $F(p, n-p-1)$ distribution.

As in simple linear regression, the following relationships hold

$$\begin{aligned} SST &= SSM + SSE \\ DFT &= DFM + DFE \end{aligned}$$

```
   Source |       SS       df       MS              Number of obs =      26
----------+------------------------------           F(  4,    21) =  479.10
    Model |  176777.061     4  44194.2653           Prob > F      =  0.0000
 Residual |  1937.13655    21  92.2445975           R-squared     =  0.9892
----------+------------------------------           Adj R-squared =  0.9871
    Total |  178714.198    25   7148.5679           Root MSE      =  9.6044

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   promotion |    1.807064   1.081039     1.67   0.109    -.4410806    4.055208
    accounts |    3.317833   .1628917    20.37   0.000     2.979082    3.656585
      brands |   -21.18498   .7879389   -26.89   0.000    -22.82359   -19.54638
   potential |    .3245121   .4677644     0.69   0.495    -.6482572    1.297281
       _cons |    178.3203   12.96032    13.76   0.000     151.3679    205.2728
------------------------------------------------------------------------------
```

Note that this F test doesn't tell you which variables are statistically significant, just whether some of them are or not.

To figure out which of the variables are most likely to be the important ones, you need to do further analysis.

The Total line in the ANOVA table describes the fit for the Reduced (H_0) model.

The Error line describes the fit for the Full (H_A) model.

The Model line describes the improvement in the fit of the Full model over the Reduced model.

F tables (Table E in book)

Give critical values for $F(df_1, df_2)$ distributions

Columns correspond to df_1 which is Model degrees of freedom

Rows correspond to df_2 which is the Error degrees of freedom

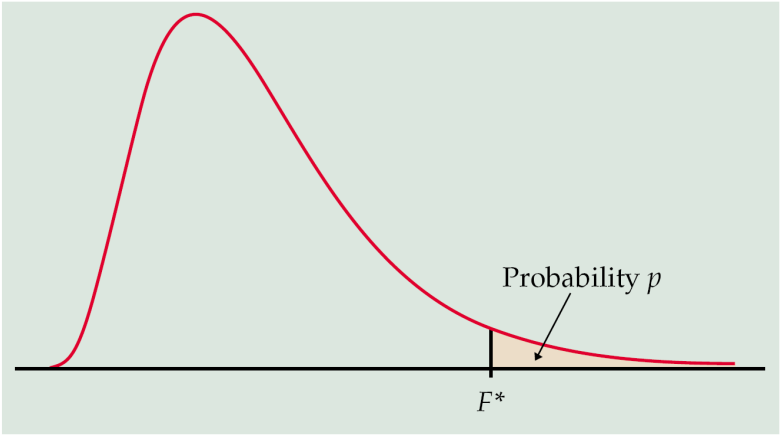


Table entry for p is the critical value F^* with probability p lying to its right.

TABLE E *F* critical values

denominator		Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28
	.010	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
	.001	405284	500000	540379	562500	576405	585937	592873	598144	602284
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86

TABLE E *F* critical values (*continued*)

		Degrees of freedom in the numerator								
	p	1	2	3	4	5	6	7	8	9
	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
18	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.001	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56
	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
19	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.001	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39
	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
20	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
21	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
ıator	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11

So for the Shingle example, if we want to do a 1% test on whether any of the predictors are useful, $F^* = 4.37$ (df $= 4, 21$). Since $F = 479.10 > 4.37$ we want to reject H_0 and conclude that some of the predictors are useful.

We can also use the table to bound p-values. Suppose we did an F test and got $F_{obs} = 3.78$ and df = 5,18. Then

$$0.01 < p\text{−value} < 0.025$$

as 3.38 $< F_{obs} <$ 4.25 (the 0.025 and 0.01 critical values).

You do not need to double p-value with the F table.

Squared Multiple Correlation R^2

$$R^2 = \frac{SSM}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

As in the simple linear regression case, it gives the proportion of the variability in the response variable described by the set of explanatory variables $x_1, x_2, \ldots, x_p$.

In the example, $R^2 = 0.9892$.

These 4 variables do a very good job of explaining the variability in the shingle sales.

As with simple linear regression, $R = \sqrt{R^2}$ is a correlation.

In this case it the correlation between y_i and $\hat{y}_i$.

Confidence Intervals for a Mean Response

Interested in the mean response of y when $x_1 = x_1^*, x_2 = x_2^*, \ldots, x_p = x_p^*$

$$\mu_y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^*$$

Estimate this with

$$\hat{\mu}_y = b_0 + b_1 x_1^* + b_2 x_2^* + \ldots + b_p x_p^*$$

The confidence interval for μ_y is

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$$

where t^* is based on a $t(n-p-1)$ distribution.

The standard error of $\hat{\mu}_y$ depends on a number of factors

- x^*

  Its smallest when $x_1^* = \bar{x}_1, x_2^* = \bar{x}_2, \ldots, x_p^* = \bar{x}_p$, and increases as the location of interest moves away from the point of means.

- s

  It proportion to s, the standard deviation of the residuals.

- Correlation of the xs.

Prediction Intervals for a Future Observation

Interested in a new observation of y when $x_1 = x_1^*, x_2 = x_2^*, \ldots, x_p = x_p^*$

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \ldots + \beta_p x_p^* + \epsilon$$

Estimate this with

$$\hat{y} = b_0 + b_1 x_1^* + b_2 x_2^* + \ldots + b_p x_p^*$$

The confidence interval for μ_y is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

where t^* is based on a $t(n - p - 1)$ distribution.

The standard error of prediction of $\hat{y}$ is

$$SE_{\hat{y}} = \sqrt{s^2 + SE_{\hat{\mu}_y}^2}$$

$SE_{\hat{y}}$ again accounts for two piece of uncertainty.

- Uncertainty about the regression surface at x^*.
- Deviations of observations from the true regression surface

Notice that $SE_{\hat{y}} \geq SE_{\hat{\mu}_y}$ and $SE_{\hat{y}} \geq s$

The statements about the magnitude of $SE_{\hat{\mu}_y}$ also hold for $SE_{\hat{y}}$.