

## Case Study: Spartina Biomass

(Data from Dr. Rick Linthurst, Ph.D. thesis, NCSU, 1979)

*Spartina alterniflora*: a marsh grass that grows in the Cape Fear Estuary of North Carolina

Data from random sampling, stratified according to two factors:

- Three locations sampled: Oak Island (OI), Smith Island (SI), Snows Marsh (SM)
- Three types of areas: revegetated “dead” areas (DVEG), “short” Spartina areas (SHRT), “tall” Spartina areas (TALL)

During the month of September, five random sites were sampled at each location-area type combination, for a total of  $n = 45$  observations.

*For now we will ignore the stratification by location and area type — these could be treated as class variables (Chapter 9) if they were of interest in the study.*

Variables observed:

- $Y$ : aerial (above-ground/water) **biomass** of Spartina ( $\text{g}/\text{m}^2$ )
- $X_1$ : **salinity** (parts per thousand)
- $X_2$ : **pH** (acidity) in water
- $X_3$ : **K** (potassium) (parts per million)
- $X_4$ : **Na** (sodium) (parts per million)
- $X_5$ : **Zn** (zinc) (parts per million)

(Note: We will work with only a subset of the data from the study. See textbook.)

Contents of file `linthurst.dat`:

1	OI	DVEG	676	33	5.00	1441.67	35184.5	16.4524
2	OI	DVEG	516	35	4.75	1299.19	28170.4	13.9852
3	OI	DVEG	1052	32	4.20	1154.27	26455.0	15.3276
4	OI	DVEG	868	30	4.40	1045.15	25072.9	17.3128
5	OI	DVEG	1008	33	5.55	521.62	31664.2	22.3312
6	OI	SHRT	436	33	5.05	1273.02	25491.7	12.2778
7	OI	SHRT	544	36	4.25	1346.35	20877.3	17.8225
8	OI	SHRT	680	30	4.45	1253.88	25621.3	14.3516
9	OI	SHRT	640	38	4.75	1242.65	27587.3	13.6826
10	OI	SHRT	492	30	4.60	1281.95	26511.7	11.7566
11	OI	TALL	984	30	4.10	553.69	7886.5	9.8820
12	OI	TALL	1400	37	3.45	494.74	14596.0	16.6752
13	OI	TALL	1276	33	3.45	525.97	9826.8	12.3730
14	OI	TALL	1736	36	4.10	571.14	11978.4	9.4058
15	OI	TALL	1004	30	3.50	408.64	10368.6	14.9302
16	SI	DVEG	396	30	3.25	646.65	17307.4	31.2865
17	SI	DVEG	352	27	3.35	514.03	12822.0	30.1652
18	SI	DVEG	328	29	3.20	350.73	8582.6	28.5901
19	SI	DVEG	392	34	3.35	496.29	12369.5	19.8795
20	SI	DVEG	236	36	3.30	580.92	14731.9	18.5056
21	SI	SHRT	392	30	3.25	535.82	15060.6	22.1344
22	SI	SHRT	268	28	3.25	490.34	11056.3	28.6101
23	SI	SHRT	252	31	3.20	552.39	8118.9	23.1908
24	SI	SHRT	236	31	3.20	661.32	13009.5	24.6917
25	SI	SHRT	340	35	3.35	672.15	15003.7	22.6758
26	SI	TALL	2436	29	7.10	528.65	10225.0	0.3729
27	SI	TALL	2216	35	7.35	563.13	8024.2	0.2703
28	SI	TALL	2096	35	7.45	497.96	10393.0	0.3205
29	SI	TALL	1660	30	7.45	458.38	8711.6	0.2648
30	SI	TALL	2272	30	7.40	498.25	10239.6	0.2105
31	SM	DVEG	824	26	4.85	936.26	20436.0	18.9875
32	SM	DVEG	1196	29	4.60	894.79	12519.9	20.9687
33	SM	DVEG	1960	25	5.20	941.36	18979.0	23.9841

34	SM	DVEG	2080	26	4.75	1038.79	22986.1	19.9727
35	SM	DVEG	1764	26	5.20	898.05	11704.5	21.3864
36	SM	SHRT	412	25	4.55	989.87	17721.0	23.7063
37	SM	SHRT	416	26	3.95	951.28	16485.2	30.5589
38	SM	SHRT	504	26	3.70	939.83	17101.3	26.8415
39	SM	SHRT	492	27	3.75	925.42	17849.0	27.7292
40	SM	SHRT	636	27	4.15	954.11	16949.6	21.5699
41	SM	TALL	1756	24	5.60	720.72	11344.6	19.6531
42	SM	TALL	1232	27	5.35	782.09	14752.4	20.3295
43	SM	TALL	1400	26	5.50	773.30	13649.8	19.5880
44	SM	TALL	1620	28	5.50	829.26	14533.0	20.1328
45	SM	TALL	1560	28	5.40	856.96	16892.2	19.2420

Columns are: observation number, location, area type, biomass, salinity, pH, K, Na, and Zn.

The data file is formatted as if it were instream data (i.e. as after a `datalines` statement), so SAS® can easily read it during a `DATA` step:

```
data linthurst (drop=obsnum);
infile 'linthurst.dat';
input obsnum loc $ type $ biomass salinity pH K Na Zn;
run;
```

A SAS® data set called `linthurst` is created, having the variables `loc`, `type`, `biomass`, `salinity`, `pH`, `K`, `Na`, and `Zn`. (`obsnum` is dropped.)

Printing the data set verifies that it was read correctly:

```
proc print;
run;
```

which results in

Obs	loc	type	biomass	salinity	pH	K	Na	Zn
1	OI	DVEG	676	33	5.00	1441.67	35184.5	16.4524
2	OI	DVEG	516	35	4.75	1299.19	28170.4	13.9852
3	OI	DVEG	1052	32	4.20	1154.27	26455.0	15.3276
4	OI	DVEG	868	30	4.40	1045.15	25072.9	17.3128
...								
45	SM	TALL	1560	28	5.40	856.96	16892.2	19.2420

(output abridged)

## Full Model Regression Analysis

For demonstration purposes we make all of the usual assumptions of linear regression:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

with  $\epsilon_i \sim \text{i.i.d.} N(0, \sigma^2)$  for  $i = 1, \dots, 45$

To anticipate results of the analysis, we will compute the **sample correlation matrix** of all variables together. First, some review ...

Recall the **correlation** of two random variables  $X$  and  $Y$  having nonzero variances:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}}$$

Properties:

$$\text{Corr}(X, Y) \in [-1, 1]$$

$\text{Corr}(X, Y) > 0 \Rightarrow X, Y$  have same tendency

$\text{Corr}(X, Y) < 0 \Rightarrow X, Y$  have opposite tendency

$$\text{Corr}(X, X) = 1$$

$X, Y$  independent  $\Rightarrow \text{Corr}(X, Y) = 0$

Recall the **sample correlation** of  $n$  observed pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$ :

$$\hat{\rho} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

assuming the  $X$ 's are not all the same and the  $Y$ 's are not all the same.

Properties similar to correlation:

$$\hat{\rho} \in [-1, 1]$$

$\hat{\rho} = 1 \Rightarrow$  perfect increasing straight line

$\hat{\rho} = -1 \Rightarrow$  perfect decreasing straight line

If the observed pairs are independent samples from the same bivariate distribution (population), then the sample correlation is often used as an estimate of the population correlation.

Recall that the sample correlation of  $(X_1, Y_1), \dots, (X_n, Y_n)$  is closely related to the simple linear regression of the  $Y$ 's on the  $X$ 's:

$$\hat{\beta}_1 = \hat{\rho} \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

and

$$R^2 = \frac{\text{SS(Regr)}}{\text{SS(Total}_{\text{corr}})} = \hat{\rho}^2$$

Even though the  $X$ 's are fixed (nonrandom) in simple linear regression, the sample correlation of  $X$  and  $Y$  still indicates the strength and direction of the  $X$ - $Y$  relationship.

Recall also (from homework) that in (multiple) linear regression

$$\sqrt{R^2} = \text{sample correlation of } \hat{Y} \text{ and } Y,$$

i.e. the sample correlation of

$$(\hat{Y}_1, Y_1), \dots, (\hat{Y}_n, Y_n).$$

The **correlation matrix** of random vector

$$\mathbf{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix}$$

is the  $m \times m$  symmetric matrix having  $\text{Corr}(Z_i, Z_j)$  in its  $(i, j)^{\text{th}}$  position.

The **sample correlation matrix** of a collection of  $n$  observed vectors

$$\mathbf{Z}^{(1)} = \begin{bmatrix} Z_1^{(1)} \\ \vdots \\ Z_m^{(1)} \end{bmatrix}, \quad \dots, \quad \mathbf{Z}^{(n)} = \begin{bmatrix} Z_1^{(n)} \\ \vdots \\ Z_m^{(n)} \end{bmatrix}$$

is the  $m \times m$  symmetric matrix having the sample correlation of  $(Z_i^{(1)}, Z_j^{(1)}), \dots, (Z_i^{(n)}, Z_j^{(n)})$  in its  $(i, j)^{\text{th}}$  position.

A (sample or true) correlation matrix has 1's on its diagonal and numbers in  $[-1, 1]$  elsewhere.

The sample correlation matrix can be formed for any such collection of vectors, even if they are not actually independent samples from some multivariate distribution.

In multiple linear regression, the sample correlation matrix is useful for preliminary assessment of two things:

- the pairwise relationships between the independent variables (and hence the degree to which any two variables might “mask” each other in the regression)
- the strengths of the regressions of the dependent variable on each independent variable alone

For the Linthurst data, consider the collection of vectors

$$\begin{bmatrix} X_{1,1} \\ X_{1,2} \\ X_{1,3} \\ X_{1,4} \\ X_{1,5} \\ Y_1 \end{bmatrix}, \quad \dots, \quad \begin{bmatrix} X_{45,1} \\ X_{45,2} \\ X_{45,3} \\ X_{45,4} \\ X_{45,5} \\ Y_{45} \end{bmatrix}$$

and let  $\hat{\rho}$  be their sample correlation matrix.

SAS® PROC REG has an option `corr` for producing the matrix  $\hat{\rho}$ :

---

```
proc reg corr;
model biomass = salinity pH K Na Zn / ss2;
run;
```

---

(We have also given the option `ss2` to the `model` statement in order to produce the partial sums of squares for the independent variables.)

The output generated by the `corr` option is as follows:

---

The REG Procedure			
Correlation			
Variable	salinity	pH	K
salinity	1.0000	-0.0513	-0.0206
pH	-0.0513	1.0000	0.0192
K	-0.0206	0.0192	1.0000
Na	0.1623	-0.0377	0.7921
Zn	-0.4208	-0.7222	0.0736
biomass	-0.1032	0.7742	-0.2046

---

Correlation			
Variable	Na	Zn	biomass
salinity	0.1623	-0.4208	-0.1032
pH	-0.0377	-0.7222	0.7742
K	0.7921	0.0736	-0.2046
Na	1.0000	0.1170	-0.2721
Zn	0.1170	1.0000	-0.6244
biomass	-0.2721	-0.6244	1.0000

---

Notice:

- K and Na are highly positively correlated (0.7921)
- pH and Zn are highly negatively correlated (-0.7222)
- both pH and Zn are highly correlated with biomass: positively for pH (0.7742) and negatively for Zn (-0.6244)
- all other correlations are relatively modest (except possibly salinity and Zn, with correlation -0.4208)

The remaining output from the run:

---

The REG Procedure  
Model: MODEL1  
Dependent Variable: biomass

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	12984700	2596940	16.37	<.0001
Error	39	6186263	158622		
Corrected Total	44	19170963			

Root MSE	398.27394	R-Square	0.6773
Dependent Mean	1000.80000	Adj R-Sq	0.6359
Coeff Var	39.79556		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type II SS
Intercept	1	1252.27938	1234.75049	1.01	0.3167	163158
salinity	1	-30.28508	24.03135	-1.26	0.2151	251921
pH	1	305.52544	87.87869	3.48	0.0013	1917306
K	1	-0.28512	0.34843	-0.82	0.4182	106211
Na	1	-0.00867	0.01593	-0.54	0.5893	47011
Zn	1	-20.67639	15.05460	-1.37	0.1775	299209

---

## Reducing the Model

Other considerations aside, it is reasonable to drop the independent variable with  $t$ -value closest to zero (equivalently, largest  $p$ -value or smallest  $F$ -value), provided it is not significant. For the Linthurst data, this variable is Na, with  $t = -0.54$ .

Can drop a variable and refit in SAS® PROC REG using the delete and print statements:

```
delete Na;
print;
run;
```

This results in fitting the model with only salinity, pH, K, and Zn as independent variables:

---

The REG Procedure  
Model: MODEL1.1  
Dependent Variable: biomass

Notice:

- the overall regression is highly significant ( $p < .0001$ )
- the five independent variables together account for about 68% of SS(Total<sub>corr</sub>) ( $R^2 = 0.6773$ )
- only the coefficient for pH is significantly different from zero ( $p = 0.0013$ ) after adjustment for the other variables

Does this mean that all variables other than pH can be dropped from the model?

**NO!** These  $t$ -tests are equivalent to the  $F$ -tests associated with the *partial* sums of squares, hence are strongly affected by (sample) correlations between the independent variables.

This does indicate that at least one variable can be dropped, but we must drop *only one variable at a time*, and refit the model (updating all of the statistics) after each drop.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12937689	3234422	20.76	<.0001
Error	40	6233274	155832		
Corrected Total	44	19170963			

Root MSE	394.75543	R-Square	0.6749
Dependent Mean	1000.80000	Adj R-Sq	0.6423
Coeff Var	39.44399		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type II SS
Intercept	1	1505.48820	1133.69403	1.33	0.1917	274802
salinity	1	-35.94326	21.47611	-1.67	0.1020	436496
pH	1	293.86111	84.47376	3.48	0.0012	1885805
K	1	-0.43882	0.20238	-2.17	0.0361	732606
Zn	1	-23.45191	14.03989	-1.67	0.1027	434796

---

Notice:

- the regression is still highly significant overall
- $R^2$  has been reduced only slightly (to 0.6749)

- the coefficient for pH is still significantly different from zero ( $p = 0.0012$ )
- additionally, the coefficient for K has become marginally significantly different from zero ( $p = 0.0361$ )

(Recall that K and Na were highly correlated, so it is not surprising that removing Na has unmasked the effect of K.)

Also,  $SS(Res)$  has increased, but  $MS(Res)$  has decreased. (Why is this possible? Why is it not especially surprising?)

Of the remaining variables, salinity and Zn are still not significant ( $p = 0.1020$  and  $p = 0.1027$ , respectively). They are about equally non-significant, though Zn has a slightly larger  $p$ -value.

Remember, these are no longer truly valid tests of significance, because we have used the data to select the current model by dropping a variable. However, they still provide a convenient rule for selecting further variables to drop.

We choose to drop Zn:

```
delete Zn;
print;
run;
```

The resulting output:

The REG Procedure					
Model: MODEL1.2					
Dependent Variable: biomass					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12502893	4167631	25.63	<.0001
Error	41	6668070	162636		
Corrected Total	44	19170963			

Root MSE	403.28135	R-Square	0.6522
Dependent Mean	1000.80000	Adj R-Sq	0.6267
Coeff Var	40.29590		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type III SS
Intercept	1	-131.24547	582.52494	-0.23	0.8229	8255.73731
salinity	1	-12.05733	16.36921	-0.74	0.4656	88239
pH	1	410.20688	48.82724	8.40	<.0001	11478835
K	1	-0.49006	0.20437	-2.40	0.0211	935178

Notice:

- $R^2$  has been somewhat reduced (from 0.6749 to 0.6522)
- pH is even more significant ( $p < 0.0001$ )
- K is still marginally significant ( $p = 0.0211$ )
- salinity is even less significant ( $p = 0.4656$ )

Recall that pH and Zn were highly (negatively) correlated, so it is not surprising that dropping Zn changes the apparent significance of pH.

Now drop salinity:

```
delete salinity;
print;
run;
```

The result:

The REG Procedure					
Model: MODEL1.3					
Dependent Variable: biomass					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12414654	6207327	38.59	<.0001
Error	42	6756309	160865		
Corrected Total	44	19170963			

Root MSE	401.07918	R-Square	0.6476
----------	-----------	----------	--------

Dependent Mean 1000.80000 Adj R-Sq 0.6308  
 Coeff Var 40.07586

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Type II SS
Intercept	1	-506.97735	279.77138	-1.81	0.0771	528239
pH	1	412.03955	48.49753	8.50	<.0001	11611782
K	1	-0.48710	0.20321	-2.40	0.0211	924266

Both remaining variables (pH and K) are significant (at nominal  $\alpha = 0.05$ ), so model reduction stops here.

This final reduced model has  $R^2 = 0.6476$ , which is only slightly smaller than the original  $R^2 = 0.6773$ .

Notice that this final model does not contain Zn, even though Zn is highly (negatively) correlated with biomass.

(We have just followed a model selection procedure called **backward elimination**, which will be studied later, in Chapter 7, along with other alternatives.)

### Analysis of the Reduced Model

The resulting estimated regression equation based on the reduced model:

$$\hat{E}(\text{biomass}) = -507.0 + 412.0 \text{ pH} - 0.4871 \text{ K}$$

with corresponding variance estimate

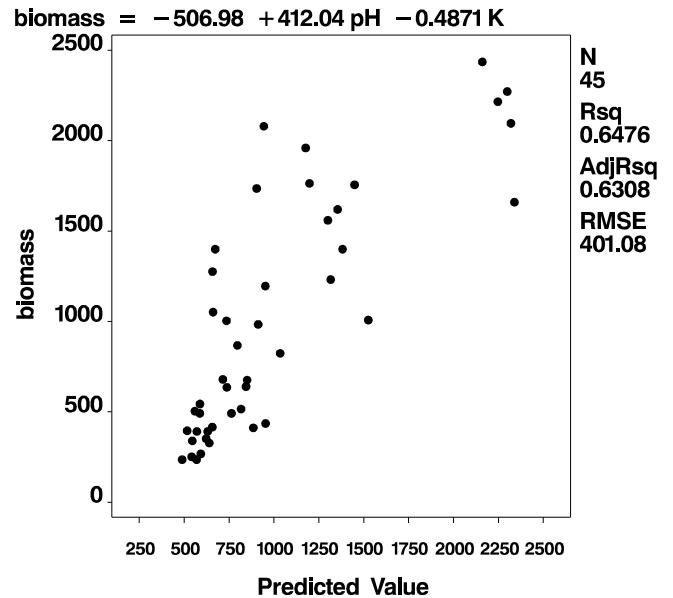
$$s^2 = 160865$$

The sample correlation of the predicted values ( $\hat{Y}$ ) and the response ( $Y$ ) is  $\sqrt{R^2} = 0.805$ . A plot of response versus predicted value ( $Y$  versus  $\hat{Y}$ ) can be obtained from PROC REG with

```
plot biomass*predicted.;
run;
```

The margins will also display the estimated regression equation and several summary statistics.

Output similar to the following:



(Exact appearance will depend on the current SAS® graphical parameter settings.)

To get the estimated variance-covariance matrix of the estimated regression coefficients from the current (reduced) model in PROC REG:

```
print covb;
run;
```

The result:

```

The REG Procedure
Model: MODEL1.3
Dependent Variable: biomass

Covariance of Estimates

Variable      Intercept      pH      K
Intercept    78272.023339   -10673.32572   -32.06557891
pH           -10673.32572   2352.0100329   -0.189496967
K            -32.06557891  -0.189496967   0.0412948105

```

Confidence intervals for the regression coefficients and observation means and prediction intervals at each observation can be obtained through options to the model statement.

The REG Procedure  
Model: MODEL2  
Dependent Variable: biomass

Sum of Residuals 0  
Sum of Squared Residuals 6756309  
Predicted Residual SS (PRESS) 7677120

We need to re-run the reduced model:

```
model biomass = pH K / clb clm cli;
run;
```

(The *PRESS* statistic is discussed in Chapter 7.)

Simultaneous confidence intervals are not directly available in PROC REG, but Bonferroni CIs can be obtained by re-running with a reduced  $\alpha$  value.

which results in (output abridged):

For example, to obtain Bonferroni 95% simultaneous confidence intervals for the three mean parameters in the reduced model, use  $\alpha = 0.05/3 \approx 0.0166667$ :

The REG Procedure  
Model: MODEL2  
Dependent Variable: biomass

...

Parameter Estimates				
Variable	DF	95% Confidence Limits		
Intercept	1	-1071.57885	57.62415	
pH	1	314.16758	509.91152	
K	1	-0.89719	-0.07700	

```
proc reg alpha=0.0166667;
model biomass = pH K / clb;
run;
```

The REG Procedure  
Model: MODEL2  
Dependent Variable: biomass

Output Statistics					
Obs	Dep Var biomass	Predicted Value	Std Error Mean Predict	95% CL Mean	
1	676.0000	850.9864	144.8405	558.6864	1143
2	516.0000	817.3781	118.2648	578.7102	1056
3	1052	661.3466	96.2405	467.1254	855.5677
4	868.0000	796.9066	78.8672	637.7461	956.0671
...					
45	1560	1301	72.1052	1155	1446

The REG Procedure  
Model: MODEL1  
Dependent Variable: biomass

...

Parameter Estimates				
Variable	DF	98.33333% Confidence Limits		
Intercept	1	-1204.63388	190.67918	
pH	1	291.10289	532.97620	
K	1	-0.99384	0.01964	

The REG Procedure  
Model: MODEL2  
Dependent Variable: biomass

Output Statistics				
Obs	95% CL Predict		Residual	
1	-9.5860	1712	-174.9864	
2	-26.4867	1661	-301.3781	
3	-171.0399	1494	390.6534	
4	-28.0040	1622	71.0934	
...				
45	478.2264	2123	259.3870	

See the textbook for nice graphs of confidence ellipsoids for these parameters. (Using SAS® to produce graphs like those would require special programming.)

## Comments: Validity and Interpretation

- Be careful interpreting confidence intervals and tests based on a model selected using the data. Because of the selection, the stated  $\alpha$  levels are no longer strictly valid.
- Based on the reduced model, may we legitimately infer that an increase in pH and/or a decrease in K (at some location) will tend to cause an increase in biomass?

NO, for at least two reasons:

- This is purely an **observational study**: it did not involve active manipulation of the independent variables by the researchers. Therefore, while you may infer *association*, you cannot infer *causality*, even if your final model is correct. The association may be only indirect (i.e. via common association with unobserved or unused variables).

This analysis does serve to identify variables that show clear natural association with biomass, but only a later phase of controlled experimentation can establish a causal relationship.

- Coefficients only reflect the change in response *when all other variables are held constant*. If actively changing one independent variable also causes other independent variables to change, the effect on the response may not be what you expect.

For example, even if artificially decreasing K really would affect biomass, if doing so also happens to decrease pH, the net effect might decrease biomass, rather than increase it.

A properly-conducted **designed experiment** can establish causality. Designed experiments will be introduced in Basic Design & Analysis of Experiments (next semester).

- May we use the reduced model equation to reliably *predict* biomass, given pH and K?

Possibly, if certain conditions are met:

- We predict for samples from the same population from which the original observations were sampled, i.e. the marshes of Cape Fear Estuary in September. (Even in this case, we must assume that the population has not been altered in some way since the sampling, e.g. by a dramatic environmental change.)
- We predict only at pairs (pH, K) that are within the range of these pairs in the data. We cannot be confident that our linear model approximation will be useful outside of this range.

Also, note that even if prediction is accurate (i.e. unbiased), it is not necessarily very precise.

Considerations in using a regression model for prediction are discussed in more detail in Chapter 7.