Multiple Linear Regression

The population model

- In a simple linear regression model, a single response measurement Y is related to a single predictor (covariate, regressor) X for each observation. The critical assumption of the model is that the conditional mean function is linear: $E(Y|X) = \alpha + \beta X$.

  In most problems, more than one predictor variable will be available. This leads to the following "multiple regression" mean function:

  $$E(Y|X) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p,$$

  where α is caled the intercept and the β_j are called slopes or coefficients.

- For example, if Y is annual income ($1000/year), X_1 is educational level (number of years of schooling), X_2 is number of years of work experience, and X_3 is gender ($X_3 = 0$ is male, $X_3 = 1$ is female), then the population mean function may be

  $$E(Y|X) = 15 + 0.8 \cdot X_1 + 0.5 \cdot X_2 - 3 \cdot X_3.$$

  Based on this mean function, we can determine the expected income for any person as long as we know his or her educational level, work experience, and gender.

  For example, according to this mean function, a female with 12 years of schooling and 10 years of work experience would expect to earn $26,600 annually. A male with 16 years of schooling and 5 years of work experience would expect to earn $30,300 annually.

- Going one step further, we can specify how the responses vary around their mean values. This leads to a model of the form

  $$Y_i = \alpha + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i.$$

  which is equivalent to writing $Y_i = E(Y|X_i) + \epsilon_i$.

  We write $X_{i,j}$ for the j^{th} predictor variable measured for the i^{th} observation.

  The main assumptions for the errors ϵ_i is that $E\epsilon_i = 0$ and $\text{var}(\epsilon_i) = \sigma^2$ (all variances are equal). Also the ϵ_i should be independent of each other.

  For small sample sizes, it is also important that the ϵ_i approximately have a normal distribution.

- For example if we have the population model

$$Y = 15 + 0.8 \cdot X_1 + 0.5 \cdot X_2 - 3 \cdot X_3 + \epsilon.$$

as above, and we know that $\sigma = 9$, we can answer questions like: "what is the probability that a female with 16 years education and no work experience will earn more than \$40,000/year?"

The mean for such a person is 24.8, so standardizing yields the probability:

$$\begin{aligned} P(Y > 40) &= P((Y - 24.8)/9 > (40 - 24.8)/9) \\ &= P(Z > 1.69) \\ &\approx 0.05. \end{aligned}$$

- Another way to interpret the mean function

$$E(Y|X) = 15 + 0.8 \cdot X_1 + 0.5 \cdot X_2 - 3 \cdot X_3.$$

is that for each additional year of schooling that you have, you can expecect to earn an additional \$800 per year, and for each additional year of work experience, you can expect to earn an additional \$500 per year.

This is a very strong assumption. For example, it may not be realistic that the gain in income when moving from from $X_2 = 20$ to $X_2 = 21$ would be equal to the gain in income when moving from $X_2 = 1$ to $X_2 = 2$.

We will discuss ways to address this later.

- The gender variable X_3 is an indicator variable, since it only takes on the values 0/1 (as opposed to X_1 and X_2 which are quantitative).

The slope of an indicator variable (i.e. β_3) is the average gain for observations possessing the characteristic measured by X_3 over observations lacking that characteristic. When the slope is negative, the negative gain is a loss.

Multiple regression in linear algebra notation

- We can pack all response values for all observations into a n-dimensional vector called the response vector:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ Y_n \end{pmatrix}$$

- We can pack all predictors into a $n \times p+1$ matrix called the design matrix:

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots X_{1p} \\ 1 & X_{21} & X_{22} & \cdots X_{2p} \\ & & \cdots & \\ & & \cdots & \\ 1 & X_{n1} & X_{n2} & \cdots X_{np} \end{pmatrix}$$

  Note the initial column of 1's. The reason for this will become clear shortly.
- We can pack the intercepts and slopes into a $p+1$-dimensional vector called the slope vector, denoted β:

$$\beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \beta_p \end{pmatrix}$$

- Finally, we can pack all the errors terms into a n-dimensional vector called the error vector:

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \cdots \\ \cdots \\ \cdots \\ \epsilon_n \end{pmatrix}$$

- Using linear algebra notation, the model

$$Y_i = \alpha + \beta_1 X_{i,1} + \cdots \beta_p X_{i,p} + \epsilon_i$$

  can be compactly written:

$$Y = X\beta + \epsilon,$$

  where $X\beta$ is the matrix-vector product.

- In order to estimate β, we take a least squares approach that is analogous to what we did in the simple linear regression case. That is, we want to minimize

$$\sum_i (Y_i - \alpha - \beta_1 X_{i,1} - \cdots \beta_p X_{i,p})^2$$

over all possible values of the intercept and slopes.

It is a fact that this is minimized by setting

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$X'X$ and $(X'X)^{-1}$ are $p+1 \times p+1$ symmetric matrices.

$X'Y$ is a $p+1$ dimensional vector.

- The fitted values are

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y,$$

and the residuals are

$$\hat{r} = Y - \hat{Y} = (I - X(X'X)^{-1}X')Y.$$

The error standard deviation is estimated as

$$\hat{\sigma} = \sqrt{\sum_i r_i^2/(n-p-1)}$$

The variances of $\hat{\alpha}$, $\hat{\beta}_1$, $\ldots$, $\hat{\beta}_p$ are the diagonal elements of the standard error matrix:

$$\hat{\sigma}^2(X'X)^{-1}.$$

- We can verify that these formulas agree with the formulas that we worked out for simple linear regression ($p = 1$). In that case, the design matrix can be written:

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdots & \cdots \\ \cdots & \cdots \\ 1 & X_n \end{pmatrix}$$

So

$$X'X = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \qquad (X'X)^{-1} = \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{pmatrix}$$

Equivalently, we can write

$$(X'X)^{-1} = \frac{1/(n-1)}{\text{var}(X)} \begin{pmatrix} \sum X_i^2/n & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix},$$

and

$$X'Y = \begin{pmatrix} \sum Y_i \\ \sum Y_i X_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ (n-1)\text{Cov}(X,Y) + n\bar{Y}\bar{X} \end{pmatrix}$$

$$(X'X)^{-1}X'Y = \begin{pmatrix} \bar{Y} - \bar{X}\text{Cov}(X,Y)/\text{Var}(X) \\ \text{Cov}(X,Y)/\text{Var}(X) \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{X} \\ \hat{\beta} \end{pmatrix}.$$

Thus we get the same values for $\hat{\alpha}$ and $\hat{\beta}$.
Moreover, from the matrix approach the standard deviations of $\hat{\alpha}$ and $\hat{\beta}$ are

$$\text{SD}(\hat{\alpha}) = \frac{\sigma\sqrt{\sum X_i^2/n}}{\sqrt{n-1}\sigma_X}$$

$$\text{SD}(\hat{\beta}) = \frac{\sigma}{\sqrt{n-1}\sigma_X},$$

which agree with what we derived earlier.

- Example: Y_i are the average maximum daily temperatures at $n = 1070$ weather stations in the U.S during March, 2001. The predictors are: latitude (X_1), longitude (X_2), and elevation (X_3).

  Here is the fitted model:

$$E(Y|X) = 101 - 2 \cdot X_1 + 0.3 \cdot X_2 - 0.003 \cdot X_3$$

  Average temperature decreases as latitude and elevation increase, but it increases as longitude increases.

  For example, when moving from Miami (latitude 25°) to Detroit (latitude 42°), an increase in latitude of 17°, according to the model average temperature decreases by $2 \cdot 17 = 34°$.

  In the actual data, Miami's temperature was 83° and Detroit's temperature was 45°, so the actual difference was 38°.

- The sum of squares of the residuals is $\sum_i r_i^2 = 25301$, so the estimate of the standard deviation of ϵ is

$$\hat{\sigma} = \sqrt{25301/1066} \approx 4.9.$$

The standard error matrix $\hat{\sigma}^2(X'X)^{-1}$ is:

$$\begin{array}{rrrr} 2.4 & -3.2 \times 10^{-2} & -1.3 \times 10^{-2} & 2.1 \times 10^{-4} \\ -3.2 \times 10^{-2} & 7.9 \times 10^{-4} & 3.3 \times 10^{-5} & -2.1 \times 10^{-6} \\ -1.3 \times 10^{-2} & 3.3 \times 10^{-5} & 1.3 \times 10^{-4} & -1.8 \times 10^{-6} \\ 2.1 \times 10^{-4} & -2.1 \times 10^{-6} & -1.8 \times 10^{-6} & 1.2 \times 10^{-7} \end{array}$$

The diagonal elements give the standard deviations of the parameter estimates, so $\mathrm{SD}(\hat{\alpha}) = 1.55$, $\mathrm{SD}(\hat{\beta}_1) = 0.03$, etc.

- One of the main goals of fitting a regression model is to determine which predictor variables are truly related to the response. This can be fomulated as a set of hypothesis tests.

For each predictor variable X_i, we may test the null hypothesis $\beta_i = 0$ against the alternative $\beta_i \neq 0$.

To obtain the p-value, first standardize the slope estimates:

$$\begin{aligned} \hat{\beta}_1/\mathrm{SD}(\hat{\beta}_1) &\approx -72 \\ \hat{\beta}_2/\mathrm{SD}(\hat{\beta}_2) &\approx 29 \\ \hat{\beta}_3/\mathrm{SD}(\hat{\beta}_3) &\approx -9 \end{aligned}$$

Then look up the result in a Z table. In this case the p-values are all extremely small, so all three predictors are significantly related to the response.

Sums of squares

- Just as with the simple linear model, the residuals and fitted values are uncorrelated:

$$\sum (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0.$$

Thus we continue to have the "SSTO = SSE + SSR" decomposition

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2.$$

- Here are the sums of squares with degrees of freedom (DF):

Source	Formula	DF
SSTO	$\sum(Y_i - \bar{Y})^2$	$n-1$
SSE	$\sum(Y_i - \hat{Y}_i)^2$	$n-p-1$
SSR	$\sum(\hat{Y}_i - \bar{Y})^2$	p

  Each mean square is a sum of squares divided by its degrees of freedom:

$$\text{MSTO} = \frac{\text{SSTO}}{n-1}, \quad \text{MSE} = \frac{\text{SSE}}{n-p-1}, \quad \text{MSR} = \frac{\text{SSR}}{p}$$

- The F statistic

$$F = \frac{\text{MSR}}{\text{MSE}}$$

  is used to test the hypothesis "all $\beta_i = 0$" against the alternative "at least one $\beta_i \neq 0$."

  Larger values of F indicate more evidence for the alternative.

  The F-statistic has $p, n-p-1$ degrees of freedom, p-values can be obtained from an F table, or from a computer program.

- **Example:** (cont.) The sums of squares, mean squares, and F statistic for the temperature analysis are given below:

Source	Sum square	DF	Mean square
Total	181439	1069	170
Error	25301	1066	24
Regression	156138	3	52046

  $F = 52046/24 \approx 2169$ on 3,1066 DF. The p-value is extremely small.

  The proportion of explained variation (PVE) is SSR/SSTO. The PVE is always between 0 and 1. Values of the PVE close to 1 indicate a closer fit to the data.

  For the temperature analysis the PVE is 0.86.

- If the sample size is large, all variables are likely to be significantly different from zero. Yet not all are equally important.

  The relative importance of the variables can be assessed based on the PVE's for various submodels:

Predictors	PVE	F
Latitude	0.75	1601
Longitude	0.10	59
Elevation	0.02	9
Longitude, Elevation	0.19	82
Latitude, Elevation	0.75	1080
Latitude, Longitude	0.85	2000
Latitude, Longitude, Elevation	0.86	1645

  Latitude is by far the most important predictor, with longitude a distant second.

Interactions

- Up to this point, each predictor variable has been incorporated into the regression function through an additive term $\beta_i X_i$. Such a term is called a main effect.

  For a main affect, a variable increases the average response by β_i for each unit increase in X_i, regardless of the levels of the other variables.

  An interaction between two variables X_i and X_j is an additive term of the form $\gamma_{ij} X_i X_j$ in the regression function.
  For example, if there are two variables, the main effects and interactions give the following regression function:

$$E(Y|X) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \gamma_{12} X_1 X_2.$$

  With an interaction, the slope of X_1 depends on the level of X_2, and vice versa. For example, holding X_2 fixed, the regression function can be written

$$E(Y|X) = (\alpha + \beta_2 X_2) + (\beta_1 + \gamma_{12} X_2) X_1,$$

  so for a given level of X_2 the response increases by $\beta_1 + \gamma_{12} X_2$ for each unit increase in X_1.

  Similarly, when holding X_1 fixed, the regression function can be written

$$E(Y|X) = (\alpha + \beta_1 X_1) + (\beta_2 + \gamma_{12} X_1) X_2,$$

  so for a given level of X_1 the response increases by $\beta_2 + \gamma_{12} X_1$ for each unit increase in X_2.

- *Example: (cont.)* For the temperature data, each of the three possible interactions was added (individually) to the model along with the three main effects. PVE's and F statistics are given below:

Interactions	PVE	F
Latitude×Longitude	0.88	1514
Latitude×Elevation	0.86	1347
Longitude×Elevation	0.88	1519

The improvements in fit (PVE) are small, nevertheless we may learn something from the coefficients.
The coefficients for the model including the latitude×longitude interaction are:

$$\begin{aligned} E(Y|X) &= 188 - 4.25\text{Latitude} + 0.61\text{Longitude} - \\ &\quad 0.003\text{Elevation} + 0.02\text{Latitude} \times \text{Longitude} \end{aligned}$$

Longitude ranges from 68° to 125° in this data set. Thus in the eastern US, the model can be aproximated as

$$E(Y|X) \approx 229 - 2.89\text{Latitude} - 0.003\text{Elevation},$$

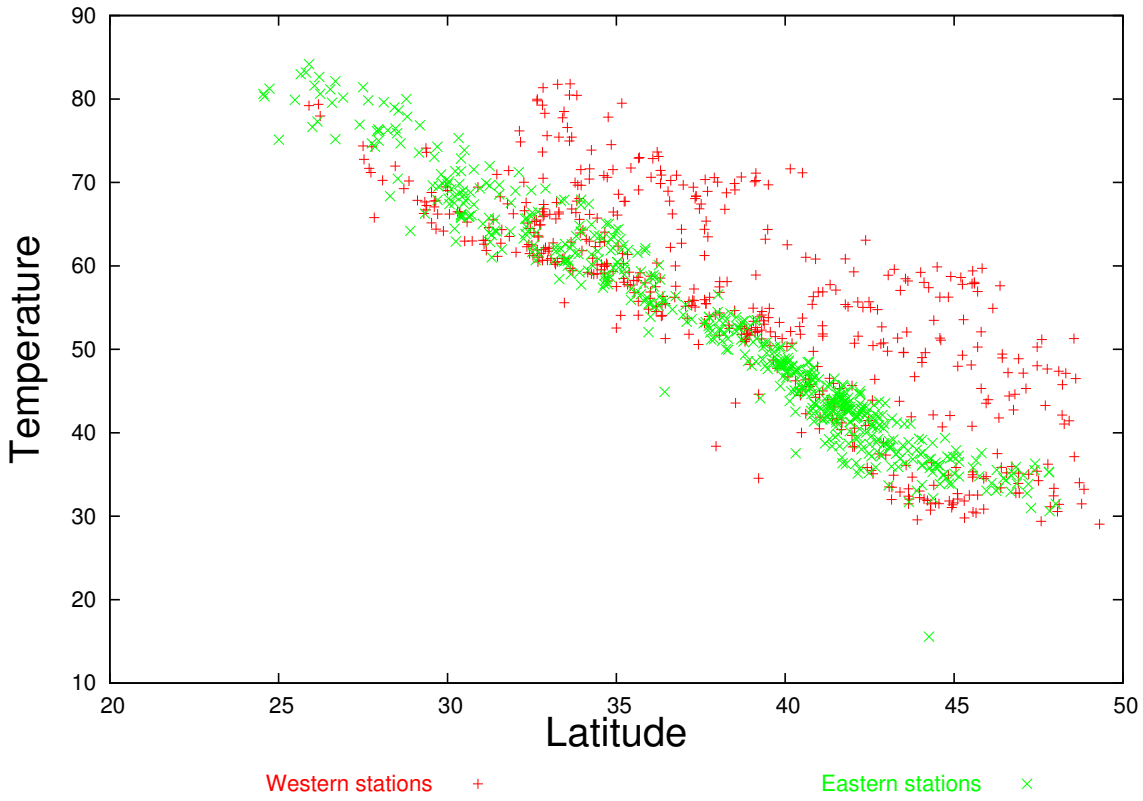
while in the western US the model can be approximated as

$$E(Y|X) \approx 264 - 1.75\text{Latitude} - 0.003\text{Elevation}.$$

This tells us that the effect of latitude was stronger in the eastern US than in the western US.

This scatterplot compares the relationships between latitude and temperature in the eastern and western US (divided at the median longitude of 93°).

The slope in the western stations is seen to be slightly closer to 0, but more notably, latitude has much less predictive power in the west compared to the east.



Polynomial regression

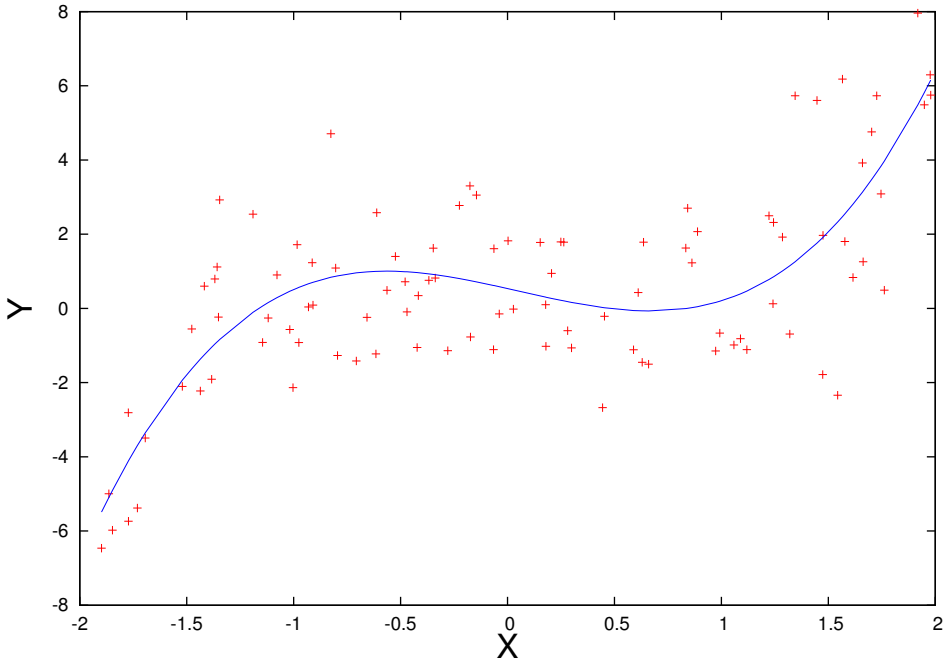
- The term "linear" in linear regression means that the regression function is linear in the coefficients α and β_j. It is not required that the X_i appear as linear terms in the regression function.

  For example, we may include power transforms of the form X_i^q for integer values of q. This allows us to fit regression functions that are polynomials in the covariates.
  For example, we could fit the following cubic model:

$$E(Y|X) = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3.$$

  This is a bivariate model, as Y and X are the only measured quantities. But we must use multiple regression to fit the model since X occurs under several power transforms.

The following data come from the population regression function $E(Y|X) = X^3 - X$, with $\text{var}(Y|X) = 4$. The fitted regression function is $\hat{E}(Y|X) = 0.54 - 1.30X - 0.18X^2 + 1.15X^3$.



- If more than one predictor is observed, we can include polynomial terms for any of the predictors.

  For example, in the temperature data we could include the three main affects along with a quadratic term for any one of the three predictors:

Quadratic term	PVE	F
Latitude	0.86	1320
Longitude	0.89	1680
Elevation	0.86	1319

  The strongest quadratic effect occurs for longitude.

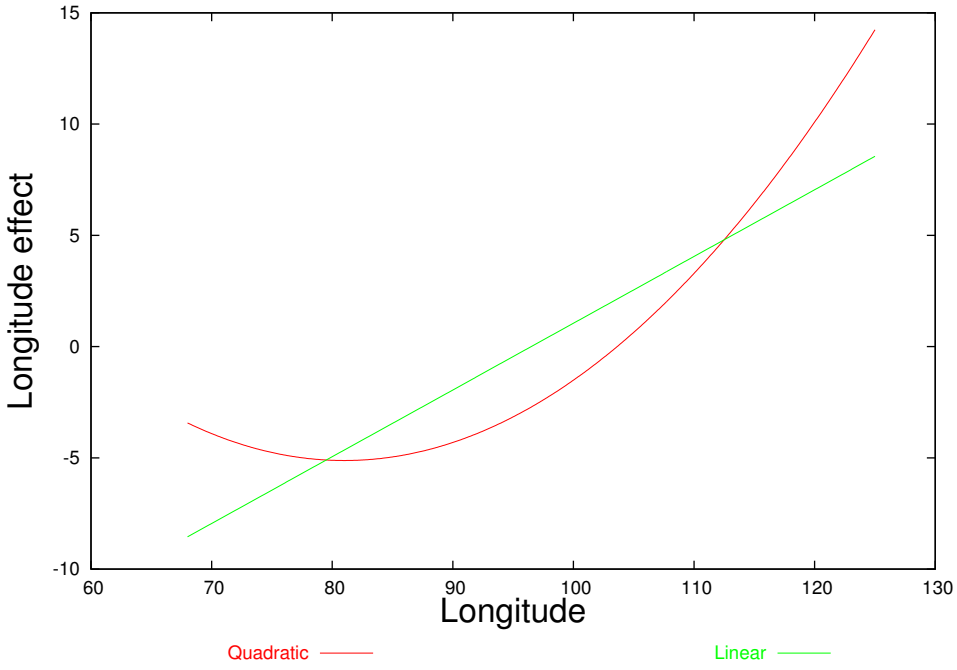
The fitted model with quadratic longitude effect is

$$\begin{aligned} E(Y|X) &= 197 - 2.09\text{Latitude} - 1.62\text{Longitude} - \\ &\quad 0.002\text{Elevation} + 0.01\text{Longitude}^2 \end{aligned}$$

Recall that a quadratic function $ax^2 + bx + c$ has a minimum if $a > 0$, a maximum if $a < 0$, and either value falls at $x = -b/2a$.

Thus the longitude effect $0.01\text{Longitude}^2 - 1.62\text{Longitude}$ has a minimum at 81°, which is around the 20^{th} percentile of our data (roughly Celevand, OH, or Columbia, SC).

The longitude effect decreases from the east coast as one moves west to around 81°, but then increases again as one continues to move further west.
This plot shows the longitude effect for the linear fit (green), and the longitude effect for the quadratic fit (red).



Model building

- Suppose you measure a response variable Y and several predictor variables $X_1, X_2, \cdots, X_p$. We can directly fit the full model

$$E(Y|X) = \alpha + \beta_1 X_1 + \cdots + \beta_p X_p,$$

but what if we are not certain that all of the variables are informative about the response?

Model building, or variable selection is the process of building a model that aims to include only the relevant predictors.

- One approach is "all subsets" regression, in which all possible models are fit (if there are p predictors then there are 2^p different models).

  A critical issue is that if more variables are included, the fit will always be better. Thus if we select the model with the highest F statistic or PVE, we will always select the full model.

  Therefore we adjust by penalizing models with many variables that don't fit much better than models with fewer variables.

  One way to do this is using the Akaike Information Criterion (AIC):

$$\text{AIC} = n \log(\text{SSE}/n) + 2(p+1).$$

Lower AIC values indicate a better model.
Here are the "all subsets" results for the temperature data:

Predictors	AIC	PVE	F
None	5499	0	0
Latitude	4016	0.75	1601
Longitude	5388	0.10	59
Elevation	5484	0.02	9
Longitude, Elevation	5281	0.19	82
Latitude, Elevation	4010	0.75	1080
Latitude, Longitude	3479	0.85	2000
Latitude, Longitude, Elevation	3397	0.86	1645

So based on the AIC we would select the full model.
As an illustration, suppose we simulate random (standard normal) "predictor variables" and include these into the temperature dataset alongside the three genuine variables.

These are the AIC PVE, and F values:

	1	10	50	100	200
AIC	3398	3399	3406	3490	3594
PVE	0.86	0.86	0.87	0.87	0.88
F	1316	474	128	64	33

The PVE continues to climb (suggesting better fit) as meaningless variables are added. The AIC increases (suggesting worse fit).

- If p is large, then it is not practical to investigate all 2^p distinct submodels. In this case we can apply forward selection.

  First find the best one-variable model based on AIC:

Predictors	AIC	PVE	F
Latitude	4016	0.75	1601
Longitude	5388	0.10	59
Elevation	5484	0.02	9

  The best model includes latitude only.
  Then select the best two variable model, where one of the variables must be latitude:

Predictors	AIC	PVE	F
Latitude, Elevation	4010	0.75	1080
Latitude, Longitude	3479	0.85	2000

  The best two-variable model includes latitude and longitude.

  If this model has worse (higher) AIC than the one-variable model, stop here. Otherwise continue to a three variable model.
  There is only one three-variable model

Predictors	AIC	PVE	F
Latitude, Longitude, Elevation	3397	0.86	1645

  Since this has lower AIC than the best two-variable model, this is our final model.

  Note that in order to arrive at this model, we never considered the longitude and elevation model.

  In general, around $p^2/2$ models most be checked in forward selection. For large p, this is far less than the 2^p models that must be checked for the all subsets aproach (i.e. if $p = 10$ then $p^2/2 = 50$ while $2^{10} = 1024$).

- A similar idea is backward selection. Start with the full model

Predictors	AIC	PVE	F
Latitude, Longitude, Elevation	3397	0.86	1645

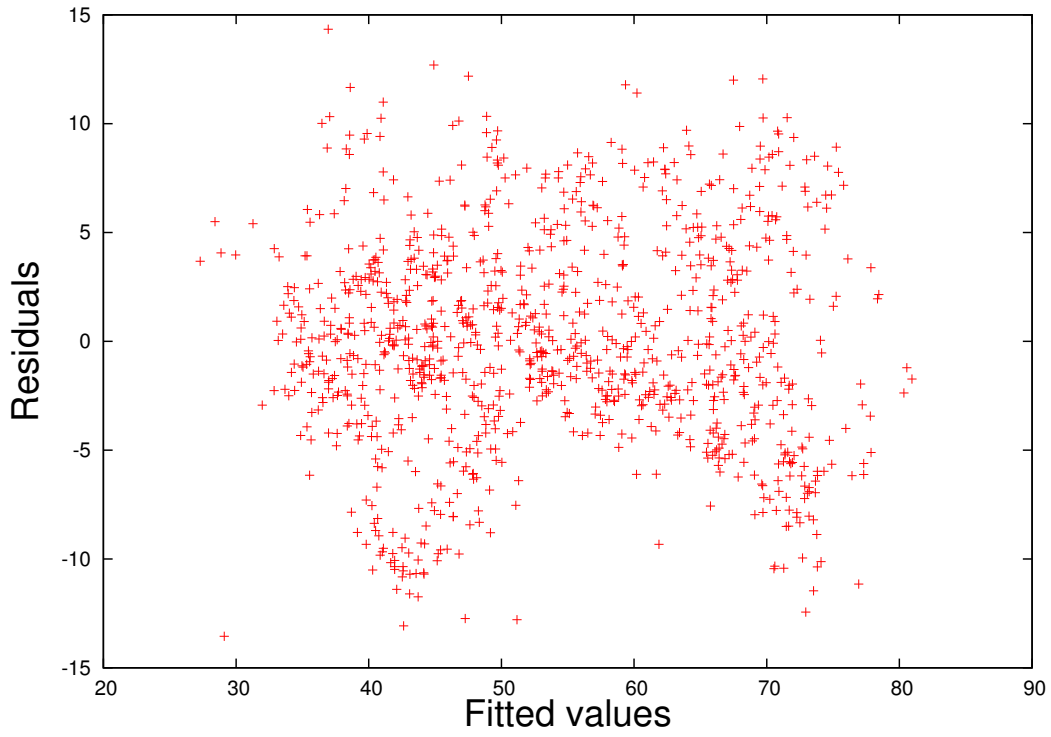
Then consider all models obtained by dropping one variable:

Predictors	AIC	PVE	F
Longitude, Elevation	5281	0.19	82
Latitude, Elevation	4010	0.75	1080
Latitude, Longitude	3479	0.85	2000

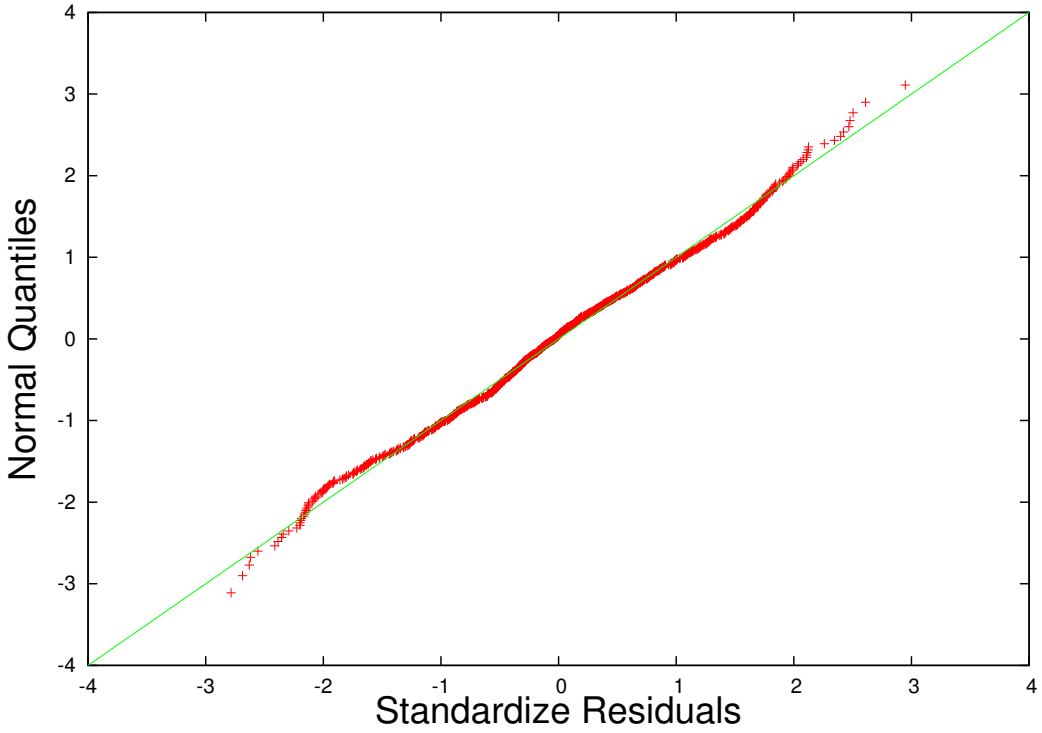
The best of these is the latitude and longitude model. Since it has higher AIC than the full model, we stop here and use the full model as our final model. If one of the two-variable models had lower AIC than the full model, then we would continue by looking at one-variable models.

Diagnostics

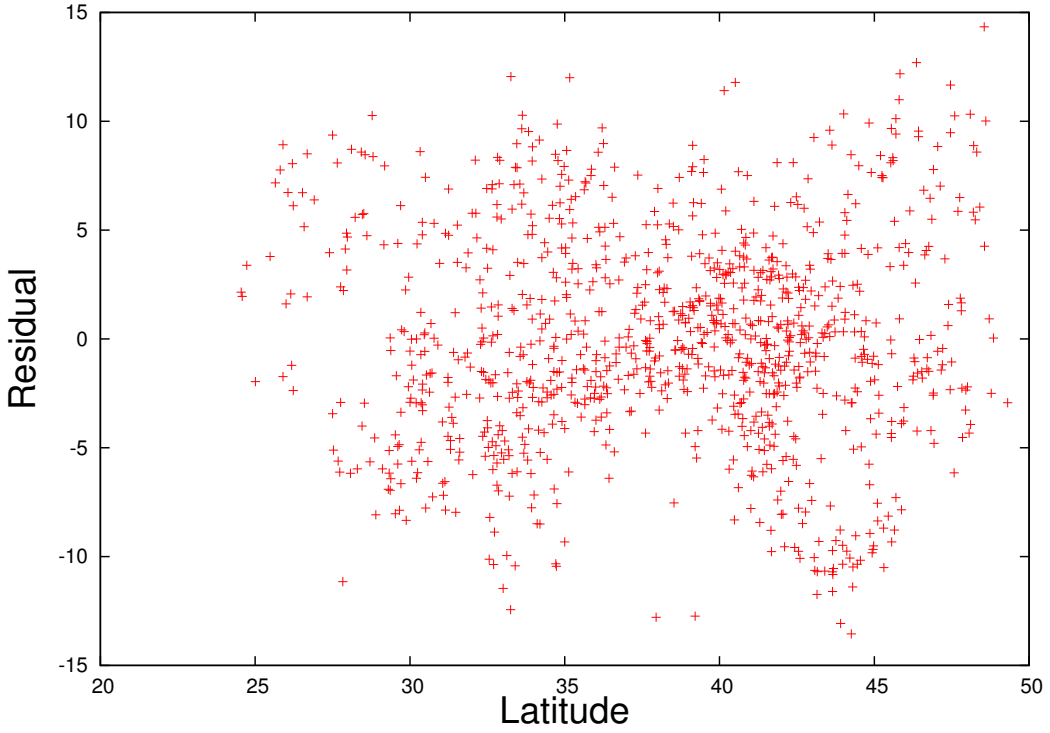
- The residuals on fitted values plot should show no pattern:



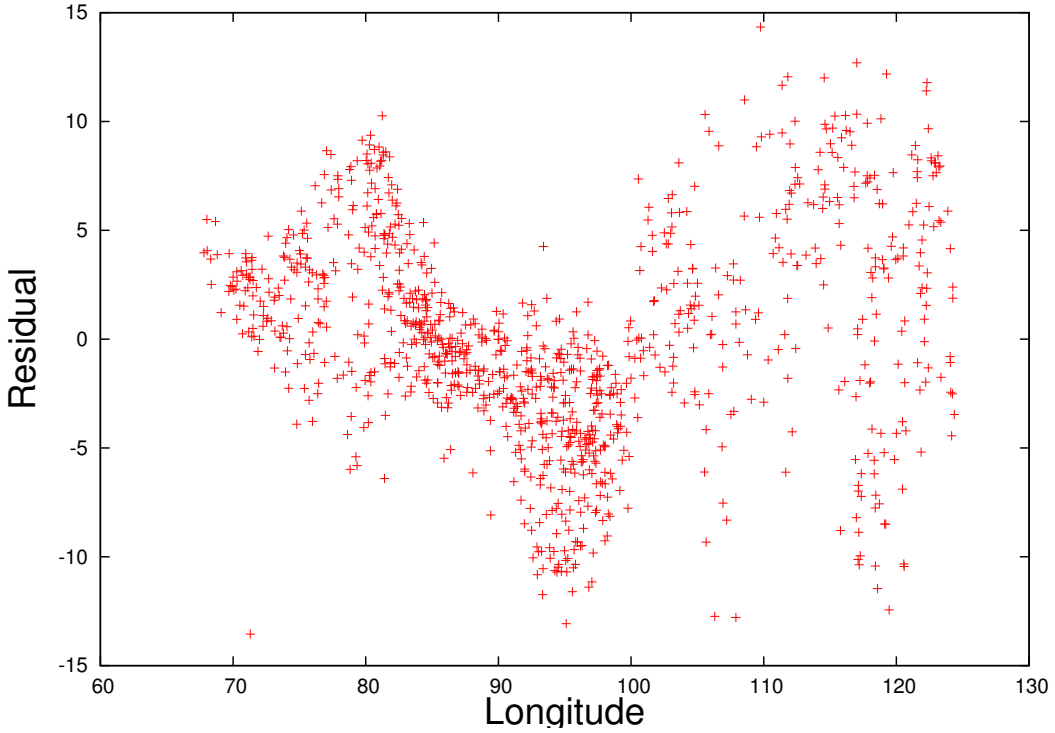
- The standardized residuals should be approxiately normal:

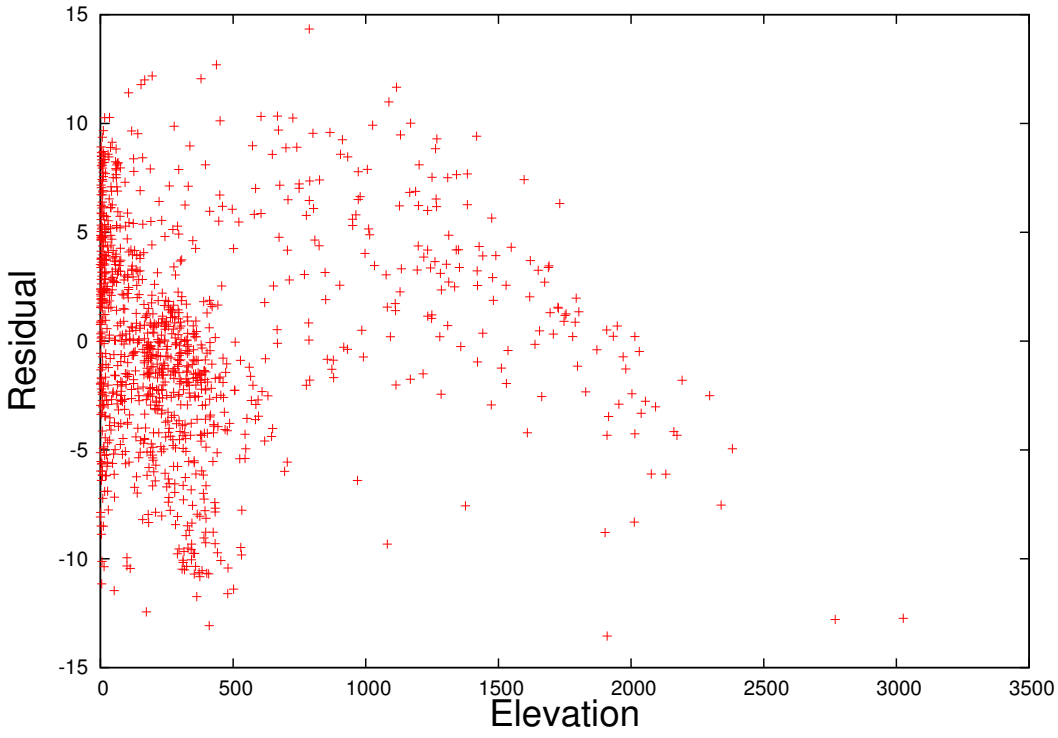


- There should be no pattern when plotting residuals against each predictor variable:



A strong suggestion that the longitude effect is quadratic:





Since two of the predictors are map coordinates, we can check whether large residuals cluster regionally: