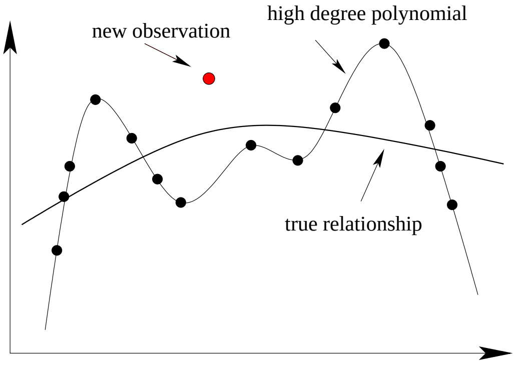
Validation of Regression Models

Recall, that the goal in most regression models is to represent some (theoretical, true) relationship between the predictor(s) and the response. If we succeed and fit a good model, then that model can be used for prediction, data description, parameter estimation or simply to gain a deeper understanding about the relationships of the variables involved. But how can you be sure that the model you fit based on the data you have is truly representative of this theoretical model? Realistically, you can't (at least not 100%). But there are ways in which you can convince yourself (and the people that hired you to do their data analysis) that your model is actually a valid representation of the true relationship.

Model validation - the techniques to make sure that a fitted regression model is reflective of the "truth" - is different from model adequacy checking (chapter 4). While model adequacy is concerned with linearity of the relationship and the residual assumptions being satisfied, model validation is conceptually different.

Example: Consider the case of fitting a polynomial regression model to a relatively small data set. Using a polynomial function of high degree will likely fit the data well and produce nice residuals. But this function may be entirely useless for new predictions, since there is no guarantee that additional data would follow the same (complicated) model.



Model validation includes testing the model in the environment it is intended to be used in. If the purpose of the model is prediction, for instance, then we should use the model to make some predictions for values we already know (but that were not used to fit the model) to see how well the model does. Users of regression models will often try to draw conclusions from the signs and magnitude of individual regression parameters (even though we know that the signs of regression parameters, for instance, can depend on all the other predictors in the model). Thus, model validation can entail a study of the regression parameters (their signs and magnitudes) to make sure they are appropriate in the context they are going to be used in.

Validation Techniques

There are three fundamentally different ways in which one can think about validating a regression model:

1) Analyzing the estimated model coefficients and predicted regression equation and comparing them with prior experience, and theoretical knowledge from the field that the data has been taken from.

2) Collection of new data (that was not used to build the model) with which to investigate the model's predictive performance.

3) Data splitting: reserving portions of the data, that are not used to fit the model and using them to check the model's predictive power on the reserved data. This process can be repeated with different splits of the data.

Analysis of Model Coefficients and Predicted Values

Ideally, the estimated coefficients of the final regression model should be stable. That means that they should remain (almost) unchanged if small changes are made on the data. These small changes could, for instance be removal of single, random data values in a large data set. Further, coefficients should have reasonable signs and magnitudes. If coefficients have counterintuitive signs compared to the known theory of the subject area, then the reason is often a model misspecification. Maybe there is multicollinearity between the predictors in the model. Or maybe the functional relationship between regressor(s) and response is misspecified. Variance inflation factors can give important clues about the stability of individual model coefficients.

Nonsensical predictions can also give us clues about a poorly validated models. If quantities that must be positive (age, weight, distance etc.) are predicted to be negative, for instance, it tells us that the model is not good for predictions in the range it is currently being used in. Here, we should differentiate between predictions within the space of observed regressor values (interpolation) and outside that range (extrapolation). The latter should only be done with caution and will likely not lead to very trustworthy predictions.

Example: The Hald Cement Data

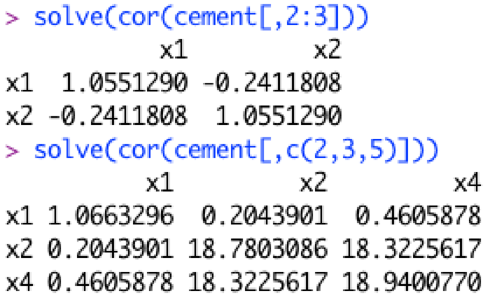
Recall, that our two final fitted models for the Cement data set were

$$\text{Model 1:} \quad \hat{y} = 52.58 + 1.468x_1 + 0.662x_2$$

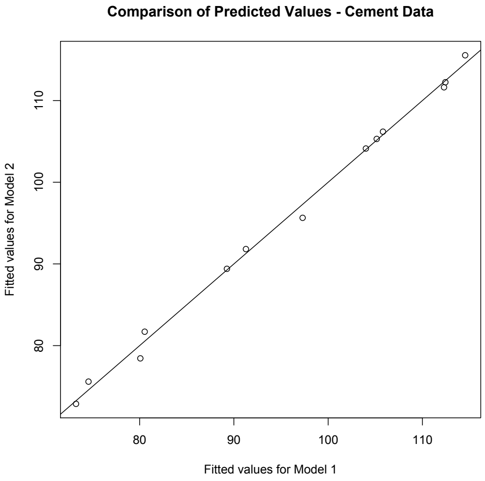
and

$$\text{Model 2:} \quad \hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$$

The regression coefficients of the two models are rather similar. So are, upon further inspection the predictions produced by the two models.



```
> solve(cor(cement[,2:3]))
           x1         x2
x1  1.0551290 -0.2411808
x2 -0.2411808  1.0551290
> solve(cor(cement[,c(2,3,5)]))
          x1         x2         x4
x1 1.0663296  0.2043901  0.4605878
x2 0.2043901 18.7803086 18.3225617
x4 0.4605878 18.3225617 18.9400770
```



Model 2 has higher variance inflation factors (VIF) than model 1, especially for parameters x_2 and x_4, since these two predictors are highly negatively correlated ($r(x_2, x_4) = -0.973$). Thus, model 1 is still the preferable model.

Model Validation Through Fresh Data - Confirmation Runs

The most effective way of model validation is to collect independent, new data after the model has been fit to an existing data set. If the existing model gives realistic predictions for the new data, users can have confidence in the validity of the model. Comparing the predictions of the model with newly collected data is sometimes called performing confirmation runs. There should be a good number ($> 10-15$) of new data points at which to test the model and they should ideally be spaced out in predictor space.

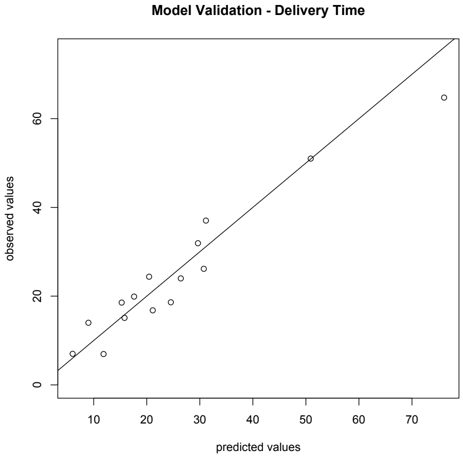
Example: The Delivery Time Data

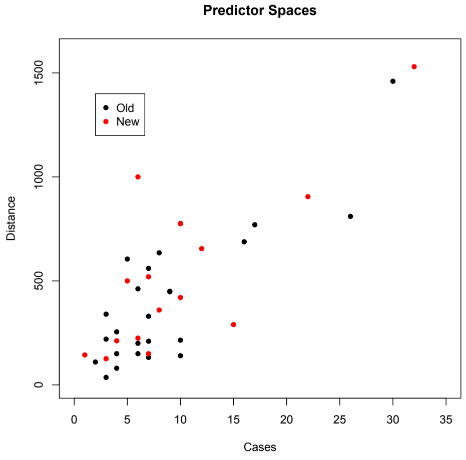
Recall the Delivery Time Data first introduced in Chapter 3. The service time for soft drink machines was predicted as a function of the number of cases stocked and the walking distance to the machines. We originally fit this model based on data from 25 machines. The data we used came from soft drink machines in Austin, San Diego, Boston and Minneapolis. There is more data (15 new observations) from Austin, Boston, San Diego and Louisville. The complete data is available in the file “DeliveryTime_2.txt”. Here, observations 1-25 are the observations we originally used to fit the model and observations 26-40 are the “new” observations.

The model that we fit for the Delivery Time data in Chapter 3 was (with y = time, x_1 = cases, x_2 = distance)

$$\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$$

Now we can use this model to predict the delivery times for the additional machines on which we have data and compare our model's prediction $\hat{y}$ with the actual observed responses for the new observations.





Most predictions are fairly accurate. The largest residual is that for the largest predicted observation (Obs 40 with $y - \hat{y} = -11.31.$). This observation is on the edge of the predictor space for the original model, thus it is possible that our model does not extrapolate very well for data outside the original range. The mean squared prediction error is

$$\text{MSP} = \frac{\sum_{i=26}^{40} (y_i - \hat{y}_i)^2}{15} = \frac{330.34}{15} = 22.023$$

That is larger than the MS_{Res} of our original model of 10.6, but that was to be expected. But the order of magnitude by which the prediction error exceeds the MS_{Res} is reasonable, so our model is fairly successful at making predictions.

Data Splitting

In many situations, collecting "fresh" data will not be possible. Maybe you're working with data that is already quite old and the situation from when the data was originally collected may have changed. Or there is simply no more money for collecting more information. In this case, you can create a validation set of data yourself, by artificially splitting the data into two parts: the estimation data and the validation data. Only the former subset will be used to fit the model, and the latter subset will be used to check the fit of the model. The "split-and-check" process is often repeated and results are averaged. This process is sometimes referred to as cross validation.

How the data should be split and how many observations should go onto each subset depends on the application that you're working with. The "leave-one-out" statistics such as the deleted or PRESS residuals that we have previously discussed are actually a form of model validation. In this case, one observation is left out (at a time) and predicted through the model based on the remaining observations.

If data has been collected longitudinally (over time) or spatially (over some set of coordinates, such as different areas of the country), then one may want to split the data with respect to time or space, respectively. For data that covers a period of several years, one could decide to leave out all the data for one of those years and to use it for validation. If the nature of the relationship between the predictor(s) and the response depends on time or space, however, then this method of selecting the validation set is likely to produce poor results. In this case, a random split of the data may be preferable. In a random split, on the other hand, you cannot guarantee that your estimation and prediction set will have the same (or even remotely similar) predictor spaces. If the predictor spaces are very different, then we cannot expect good predictions, either. If, on the other hand, the validation set is a true subset of the estimation set in the sense of predictor space, we won't be able to judge how well the model extrapolates.

One obvious solution to the problems described above is to repeat the validation analysis many times for different splits of the data. In fact, bootstrap techniques which we might briefly discuss later are an extension of this idea.

There are some theoretical approaches to splitting data sets that aim to maximize information gained from the split based on the predictor and response values of the actual data. They aim to make the volume of the predictor space that is spanned by the estimation and validation sets, respectively, as similar as possible.

General Considerations when splitting data for estimation and validation:

1) Some data sets are too small for effective splits. We need a reasonable error degree of freedom for the fitted model $(n-p)$. Rule of thumb: only split data sets of size $n \geq 2p+25$. Then estimation and validation sets can have roughly equal size.

2) Theoretically, the ratio of estimation set sample size and validation set sample size can be anything. Usually, the estimation set is larger than the validation set. The validation set should have at least about 10-15 data points in it.

3) If there are points in the data set that are "almost" replicates, then these points should *not* be split into the estimation and validation set. Instead, near neighbors could be combined into single data points.

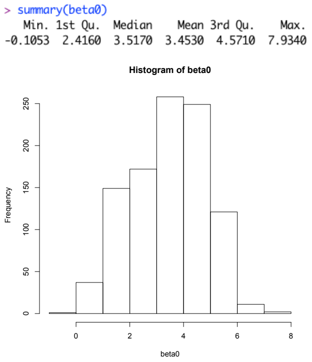
4) Double cross-validation splits the data into estimation and prediction sets. First, the estimation set is used to fit a model and the model is validated with the prediction set. Then the roles of the two sets are reversed and the analysis is repeated. The two models thus obtained can be compared to each other to check for stability of the estimation of the model parameters.

Example: The Delivery Time Data

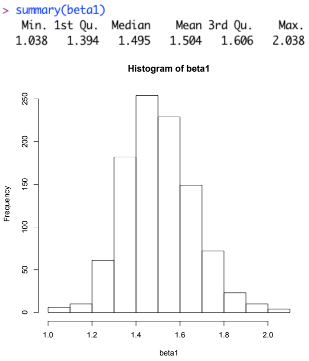
The full data set on the Delivery Time Data now contains 40 observations on Time, Cases, and Distance. Observe what happens if the data is repeatedly split into two subsets of size $n = 20$ each, of which one subset is used for estimation of the model parameters and the other subset is used to evaluate how well the model fits.

Iteration	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MSP
1	2.63213029	1.5373082	0.0179689	26.751696
2	4.15934576	1.4549023	0.0120984	14.27547
3	3.03841517	1.4977342	0.0141457	16.393103
4	4.18490871	1.4693527	0.0134895	17.656314
5	4.25211049	1.4184299	0.0149593	11.080013
⋮	⋮	⋮	⋮	⋮

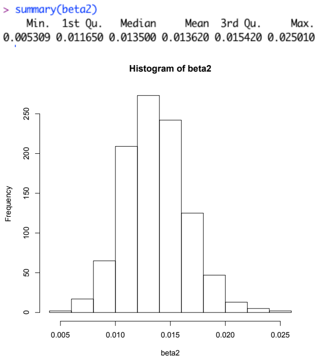
We can see that the estimate of the intercept is more variable than the estimate of both slopes. If we want to get better statistics, we have to repeat the subset selection (many) more times. For instance, for 1000 randomly selected subsets of the data we obtain the following summary statistics:



```
> summary(beta0)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1053  2.4160  3.5170  3.4530  4.5710  7.9340
```

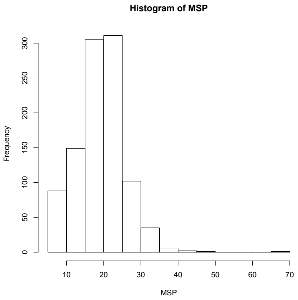


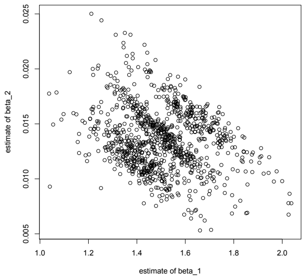
```
> summary(beta1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.038   1.394   1.495   1.504   1.606   2.038
```

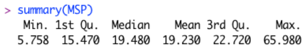


```
> summary(beta2)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.005309 0.011650 0.013500 0.013620 0.015420 0.025010
```

The distribution of the mean square error for prediction is skewed as it depends on whether or not the two extreme observations in this case are included in the estimation or in the prediction set. Not surprisingly, the estimates of the two slopes in this problem are negatively correlated.







```
> summary(MSP)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.758  15.470  19.480  19.230  22.720  65.980
```