Statistical Models in R

Some Examples

Steven Buechler

Department of Mathematics
276B Hurley Hall; 1-6233

Fall, 2007

Outline

Regression

Regression analysis is the appropriate statistical method when the response variable and all explanatory variables are continuous. Here, we only discuss linear regression, the simplest and most common form.

Remember that a statistical model attempts to approximate the response variable Y as a mathematical function of the explanatory variables $X_1, \ldots, X_n$. This mathematical function may involve parameters. Regression analysis attempts to use sample data find the parameters that produce the best model

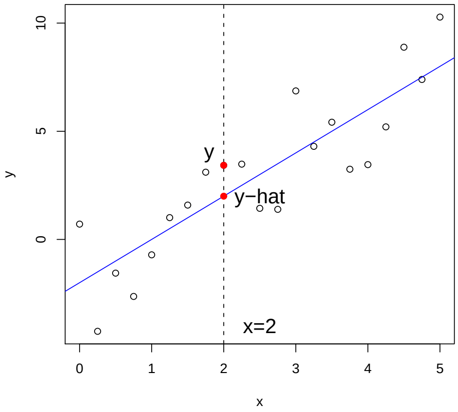
Linear Models

The simplest such model is a linear model with a unique explanatory variable, which takes the following form.

$$\hat{y} = a + bx.$$

Here, y is the response variable vector, x the explanatory variable, $\hat{y}$ is the vector of fitted values and a (intercept) and b (slope) are real numbers. Plotting y versus x, this model represents a line through the points. For a given index i, $\hat{y}_i = a + bx_i$ approximates y_i. Regression amounts to finding a and b that gives the best fit.

Linear Model with 1 Explanatory Variable



Plotting Commands

for the record

The plot was generated with test data `xR, yR` with:

```
> plot(xR, yR, xlab = "x", ylab = "y")
> abline(v = 2, lty = 2)
> abline(a = -2, b = 2, col = "blue")
> points(c(2), yR[9], pch = 16, col = "red")
> points(c(2), c(2), pch = 16, col = "red")
> text(2.5, -4, "x=2", cex = 1.5)
> text(1.8, 3.9, "y", cex = 1.5)
> text(2.5, 1.9, "y-hat", cex = 1.5)
```

Linear Regression = Minimize RSS

Least Squares Fit

In linear regression the best fit is found by minimizing

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2.$$

This is a Calculus I problem. There is a unique minimum and unique a and b achieving the minimum.

No Difference with Multiple

Explanatory Variables

Suppose there are k continuous explanatory variables, $x_1, \ldots, x_k$. A linear model of y in these variables is of the form

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$

The multiple linear regression problem is to find a, $b_1, \ldots, b_k$ that minimze RSS. With the mild assumption of independence in the $x_1, \ldots, x_k$, there is a again a unique solution. It's just a problem in matrix algebra rather than high school algebra.

Is the Model Predictive?

No assumptions have yet been made about the distribution of y or any other statistical properties. In modeling we want to calculate a model (a and $b_1, \ldots, b_k$) from the sample data and claim the same relationship holds for other data, within a certain error. Is it reasonable to assume the linear relationship generalizes?

Given that the variables represent sample data there is some uncertainty in the coefficients. With other sample data we may get other coefficients. What is the error in estimating the coefficients?

Both of these issues can be addressed with some additional assumptions.

Is the Model Predictive?

No assumptions have yet been made about the distribution of y or any other statistical properties. In modeling we want to calculate a model (a and $b_1, \ldots, b_k$) from the sample data and claim the same relationship holds for other data, within a certain error. Is it reasonable to assume the linear relationship generalizes?

Given that the variables represent sample data there is some uncertainty in the coefficients. With other sample data we may get other coefficients. What is the error in estimating the coefficients?

Both of these issues can be addressed with some additional assumptions.

Is the Model Predictive?

No assumptions have yet been made about the distribution of y or any other statistical properties. In modeling we want to calculate a model (a and $b_1, \ldots, b_k$) from the sample data and claim the same relationship holds for other data, within a certain error. Is it reasonable to assume the linear relationship generalizes?

Given that the variables represent sample data there is some uncertainty in the coefficients. With other sample data we may get other coefficients. What is the error in estimating the coefficients?

Both of these issues can be addressed with some additional assumptions.

Assumptions in Linear Models

Given a linear model $\hat{y} = a + b_1 x_1 + \cdots + b_k x_k$ of the response variable y, the validity of the model depends on the following assumptions. Recall: the residual vector is $y - \hat{y}$.

Homoscedasticity (Constant Variance) The variance of the residuals is constant across the indices. The points should be evenly distributed around the mean. Plotting residuals versus fitted values is a good test.

Normality of Errors The residuals are normally distributed.

Assumptions in Linear Models

Given a linear model $\hat{y} = a + b_1 x_1 + \cdots + b_k x_k$ of the response variable y, the validity of the model depends on the following assumptions. Recall: the residual vector is $y - \hat{y}$.

Homoscedasticity (Constant Variance) The variance of the residuals is constant across the indices. The points should be evenly distributed around the mean. Plotting residuals versus fitted values is a good test.

Normality of Errors The residuals are normally distributed.

Assumptions in Linear Models

Given a linear model $\hat{y} = a + b_1 x_1 + \cdots + b_k x_k$ of the response variable y, the validity of the model depends on the following assumptions. Recall: the residual vector is $y - \hat{y}$.

Homoscedasticity (Constant Variance) The variance of the residuals is constant across the indices. The points should be evenly distributed around the mean. Plotting residuals versus fitted values is a good test.

Normality of Errors The residuals are normally distributed.

Assessment Methods

These conditions are verified in R linear fit models with plots, illustrated later.

If a plot of residuals versus fitted values shows a dependence pattern then a linear model is likely invalid. Try transforming the variables; e.g., fit $log(y)$ instead of y, or include more complicated explanatory variables, like x_1^2 or x_1x_2.

With normality of residuals, RSS satisfies a chi-squared distribution. This can be used as a measure of the model's quality and compare linear models with different sets of explanatory variables.

Assessment Methods

These conditions are verified in *R* linear fit models with plots, illustrated later.

If a plot of residuals versus fitted values shows a dependence pattern then a linear model is likely invalid. Try transforming the variables; e.g., fit $log(y)$ instead of y, or include more complicated explanatory variables, like x_1^2 or $x_1 x_2$.

With normality of residuals, *RSS* satisfies a chi-squared distribution. This can be used as a measure of the model's quality and compare linear models with different sets of explanatory variables.

Assessment Methods

These conditions are verified in *R* linear fit models with plots, illustrated later.

If a plot of residuals versus fitted values shows a dependence pattern then a linear model is likely invalid. Try transforming the variables; e.g., fit $log(y)$ instead of y, or include more complicated explanatory variables, like x_1^2 or $x_1 x_2$.

With normality of residuals, RSS satisfies a chi-squared distribution. This can be used as a measure of the model's quality and compare linear models with different sets of explanatory variables.

Linear Models in R

Given: A response variable `Y` and explanatory variables `X1, X2, ...,Xk` from continuous random variables.

A linear regression of `Y` on `X1, X2, ..., Xk` is executed by the following command.

```
> lmFit <- lm(Y ~ X1 + ... + Xk)
```

The values of the estimated coefficients and statistics measuring the goodness of fit are revealed through

```
summary(lmFit)
```

Example Problem

There is one response variable `yy` and five explanatory variables `x1`, `x2`, `x3`, `x4`, `x5`, all of length 20. The linear fit is executed by

```
> lmFit1 <- lm(yy ~ x1 + x2 + x3 + x4 +
+       x5)
```

Results of the Linear Fit

```
> summary(lmFit1)

  Call:
lm(formula = yy ~ x1 + x2 + x3 + x4 + x5)

Residuals:
   Min      1Q Median      3Q     Max
-1.176 -0.403 -0.106  0.524  1.154

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.660      1.098    4.24  0.00082
x1             3.235      1.207    2.68  0.01792
x2             3.147      0.688    4.57  0.00043
x3            -6.486      1.881   -3.45  0.00391
x4            -1.117      0.596   -1.87  0.08223
x5             1.931      0.241    8.03  1.3e-06
```

Results of the Linear Fit

continued

```
  (Intercept) ***
x1          *
x2          ***
x3          **
x4          .
x5          ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.684 on 14 degrees of freedom
Multiple R-Squared: 0.974,       Adjusted R-squared: 0.965
F-statistic:  106 on 5 and 14 DF,  p-value: 1.30e-10
```

What Class is lmFit1?

```
> class(lmFit1)

[1] "lm"

> names(lmFit1)

 [1] "coefficients"  "residuals"
 [3] "effects"       "rank"
 [5] "fitted.values" "assign"
 [7] "qr"            "df.residual"
 [9] "xlevels"       "call"
[11] "terms"         "model"
```

These can be used to extract individual components, e.g., `lmFit1$fitted.values` is the vector of fitted values, the "hat" vector.

Explanation of Coefficients

The `Estimate` column gives the model's estimate of a (Intercept) and `b1, b2, b3, b4, b5`. The vector of fitted values is

```
yy-hat = 4.660 + 3.235*x1 + 3.147*x2 - 6.486*x3 +
         -1.117*x4 + 1.931*x5
```

From the assumed normal distribution of the residuals it's possible to estimate the error in the coefficients (see the second column). The t test is a test of null hypothesis that the coefficient is 0. If the p-value in the fourth column is < 0.05 then the variable is significant enough to be included in the model.

Measures of Fit Quality

Several parameters are given that measure the quality of the fit. The distribution of values of the residuals is given.

The model degrees of freedom, `df`, is the length of `yy` minus the number of parameters calculated in the model. Here this is $20 - 6 = 14$. By definition the residual standard error is

$$\sqrt{\frac{RSS}{df}}.$$

Clearly, it's good when this is small.

Measures of Fit Quality

A quantity frequently reported in a model is R^2. Given the y values $y_1, \ldots, y_n$, the mean of y, $\bar{y}$, and the fitted values $\hat{y}_1, \ldots, \hat{y}_n$,

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

This is a number between 0 and 1. The quality of fit increases with R^2. The adjusted R^2 does some adjustment for degrees of freedom.

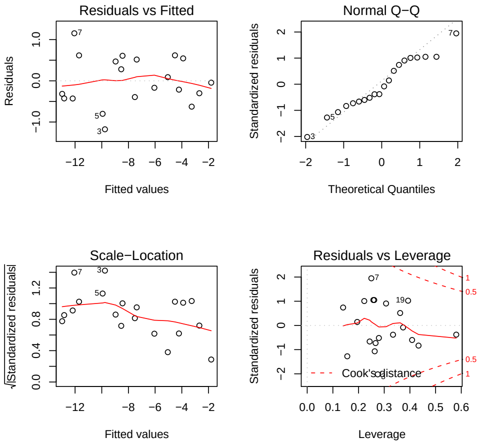
In our example R^2 is 0.974, which is very high.

Plots to Assess the Model

Remember the assumptions on the residuals needed to consider the linear model valid. We need an even scatter of residuals when plotted versus the fitted values, and a normal distribution of residuals. *R* produces 4 plots we can use to judge the model. The following code generates the 4 plots in one figure, then resets the original graphic parameters.

```
> oldpar <- par(mfrow = c(2, 2))
> plot(lmFit1)
> par(oldpar)
```

Plots of lmFit1



Using R^2 to Compare Models?

A problem with R^2, though, is that it doesn't follow a distribution. We can't compare the R^2's in two models and know when one is meaningfully better.

Just as an F statistic assessed the significance of an anova model, we use a statistic that follows an F distribution to compare two linear models, and to compare a single model to the null model.

Comparing Linear Models

A typical concern in a linear modeling problem is whether leaving a variable out meaningfully diminishes the quality of the model. There is some disadvantage in that RSS may increase some in the smaller model, however using fewer variables is a simpler model, always a plus. We need to measure the trade-off.

The F Statistic

Suppose the data contain N samples ($N =$ length of y). Consider two linear models M_0, M_1. M_0 has p_0 variables and RSS value RSS_0. Model M_1 has $p_1 > p_0$ variables, the variables in M_0 are included in those used in M_1, and the RSS value is RSS_1. Let F be the number defined as

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)}.$$

Under the assumption that the residuals are normally distributed, F satisfies an F distribution with $p_1 - p_0$ and $N - p_1 - 1$ degrees of freedom.

Tested with Anova

There is a simple way to execute this test in *R*. If `fit1` and `fit2` are the objects returned by `lm` for the two nested models, the test is executed by

```
> compMod <- anova(fit1, fit2)
```

This is not `aov`, which models a continuous variable against a factor. The similarity is that both use the F distribution to measure the statistic; all such tests are an analysis of variance in some form.

Test One Model Against the Null

In the `summary(lmFit1)` output the last line reports an F statistic. This is a comparison between the model and the null model, that sets all coefficients to 0 except the intercept. This statistic can be $> .05$ when y has no dependence on the explanatory variables.

Remove Variable from lmFit1

and Test the Result

The variable judged least significant in `lmFit1` is `x4`. For it, the p-value is .08, which is above the threshold. Generate another model without it.

```
> lmFit2 <- lm(yy ~ x1 + x2 + x3 + x5)
```

Inspect lmFit2

```
> summary(lmFit2)

  Call:
lm(formula = yy ~ x1 + x2 + x3 + x5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15346 -0.33076  0.00698  0.29063  1.30315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.622      1.024    3.54  0.00300
x1             1.013      0.237    4.27  0.00067
x2             2.137      0.461    4.63  0.00032
x3            -2.975      0.152  -19.63  4.1e-12
x5             1.935      0.260    7.44  2.1e-06
```

Inspect lmFit2

continued

```
(Intercept) **
x1          ***
x2          ***
x3          ***
x5          ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.739 on 15 degrees of freedom
Multiple R-Squared: 0.968,        Adjusted R-squared: 0.959
F-statistic:  113 on 4 and 15 DF,  p-value: 5.34e-11
```

Compare the Two Models with Anova

```
> compFit1Fit2 <- anova(lmFit2, lmFit1)
> compFit1Fit2

Analysis of Variance Table

Model 1: yy ~ x1 + x2 + x3 + x5
Model 2: yy ~ x1 + x2 + x3 + x4 + x5
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1      15 8.18
2      14 6.54  1      1.64 3.5  0.082 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use lmFit2 In Place of lmFit1

Since the p-value is 0.082, which is > 0.05, we accept the null hypothesis that the model using 5 variables (`lmFit1`) is not significantly better than the model using 4 variables (`lmFit2`). In this situation we use `lmFit2` as a model preferred over `lmFit1`.

Variable Selection

a simple but typical case

The following steps yield a model with the fewest number of variables that is statistically as meaningful as a larger model.

- Generate an initial model using all reasonable explanatory variables.
- Identify the variable with the smallest p-value.
- Compute a linear model using the smaller set of variables.
- Compute an anova for the two models. If the p-value is < 0.05 then the larger model is significantly better than the smaller. We accept the larger model as optimal. Otherwise, repeat steps 2–4.

Variable Selection

a simple but typical case

The following steps yield a model with the fewest number of variables that is statistically as meaningful as a larger model.

- Generate an initial model using all reasonable explanatory variables.
- Identify the variable with the smallest p-value.
- Compute a linear model using the smaller set of variables.
- Compute an anova for the two models. If the p-value is < 0.05 then the larger model is significantly better than the smaller. We accept the larger model as optimal. Otherwise, repeat steps 2–4.

Variable Selection

a simple but typical case

The following steps yield a model with the fewest number of variables that is statistically as meaningful as a larger model.

- Generate an initial model using all reasonable explanatory variables.
- Identify the variable with the smallest p-value.
- Compute a linear model using the smaller set of variables.
- Compute an anova for the two models. If the p-value is < 0.05 then the larger model is significantly better than the smaller. We accept the larger model as optimal. Otherwise, repeat steps 2–4.

Variable Selection

a simple but typical case

The following steps yield a model with the fewest number of variables that is statistically as meaningful as a larger model.

- Generate an initial model using all reasonable explanatory variables.
- Identify the variable with the smallest p-value.
- Compute a linear model using the smaller set of variables.
- Compute an anova for the two models. If the p-value is < 0.05 then the larger model is significantly better than the smaller. We accept the larger model as optimal. Otherwise, repeat steps 2–4.

Linear Models are Broadly Applicable

More complicated models can be generated by transforming a variable or including interactions between variables. Instead of fitting y to

$$a + b_1 x1 + b2x_2$$

it may be more meaningful to fit $log(y)$ to

$$a + c_1 x_1 + c_2 x_1^2 + c_3 x_2 + c_4 x_1 \cdot x_2.$$

This is still considered a linear model since it is linear in the parameters. R's handling of generalized lienar models is applicable here.