Goodness of Fit in Linear Regression

Basic Ideas

"Goodness of Fit" of a linear regression model attempts to get at the perhaps surprisingly tricky issue of how well a model fits a given set of data, or how well it will predict a future set of observations.

That this is a tricky issue can best be summarized by a quote from famous Bayesian statistician George Box, who said:

"All Models are wrong, but some are useful."

It may initially surprise you to see such a quote, but a bit of thought should convince you that it is true. Consider *any* linear regression model, which looks like this

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

and has various assumptions attached to it, such as *exact* linearity of all relationships, normality of residuals or errors from the model, constant residual variance throughout the range, and so on. Is it realistic to believe that *all* of these assumptions *perfectly* hold? Probably just about never!

But: The assumptions may hold "closely enough" for the model to be useful in practice, even though we are almost always very sure the model is in fact false, for one or more reasons.

The problem with "goodness of fit", then, is how to define "closely enough"? This is a concept that is hard to pin down for any particular case, let alone in general.

The difficulty is that "closely enough" must refer to a specific purpose for which the regression can be used (for example, for making predictions, or for making inferences about the effects of one variable on another), and even once this is found, one must quantitatively operationalize "closely enough".

No one has come up with a perfect measure of goodness of fit for statistical models, although there has been and continues to be much research in the area. We will look at a variety of concepts that fall into the general category of goodness of fit, including:

- Examining residuals from the model
- Outlier detection
- A global measure of “variance explained”, R^2
- A global measure of “variance explained” that is adjusted for the number of parameters in a model, adjusted R^2
- Chi-square goodness of fit tests
- Model validation via an outside data set or by splitting a data set

For each of the above, we will define the concept, see an example, and discuss the advantages and disadvantages of each.

Examining residuals from the model

We have already discussed looking at residuals from a model, and it remains one of the most informative methods by which to investigate model fit. Residuals can be used descriptively, usually by looking at histograms or scatter plots of residuals, and also form the basis for several other methods we will examine.

Recall that if we have a model like

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

then the residual for the i^{th} observation from the model is given by:

$$Y_i - \hat{Y}_i = Y_i - \left(\hat{\alpha} + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_p X_{ip}\right)$$

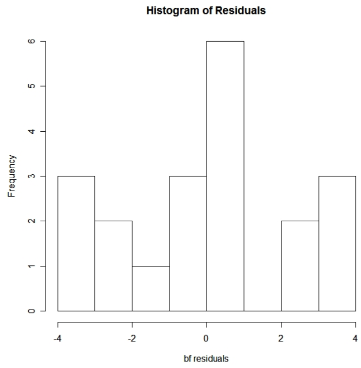
where Y_i is the observed dependent variable and the X_{ij}'s represent the observed covariate values for the i^{th} observation.

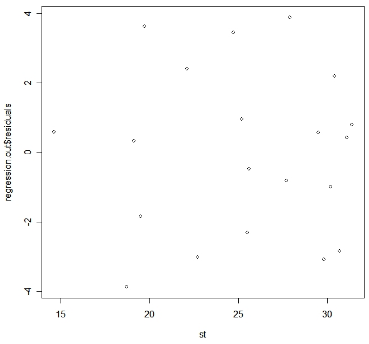
Consider again the example from last class, dealing with body fat, and let’s look at the residuals from the model that contained the covariates *st* and *mc*. Here are the histograms and residual scatter plots of interest:

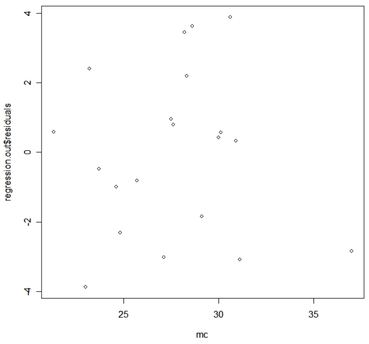
```
> regression.out <- lm(bf ~ st + mc)
```

```
> hist(regression.out$residuals, main="Histogram of Residuals", xlab = "bf residuals")
> plot(st, regression.out$residuals)

> plot(mc, regression.out$residuals)
```







The histogram is not strongly normal in shape, but, then again, there are just 20 observations. Both scatter plots look reasonable, i.e., no obvious departures from constant variance or linearity.

Conclusion from looking at the residuals: Normality may or may not be met, but probably not fatal to using the model (which is robust to non-normality to a certain extent anyway, especially if remains symmetric). Linearity and constant variance assumptions probably met. No strong reason to investigate other models.

The above conclusions are "hand-waving", we will see below some more formal tests of model fit (but we will be critical of these too).

Outlier detection

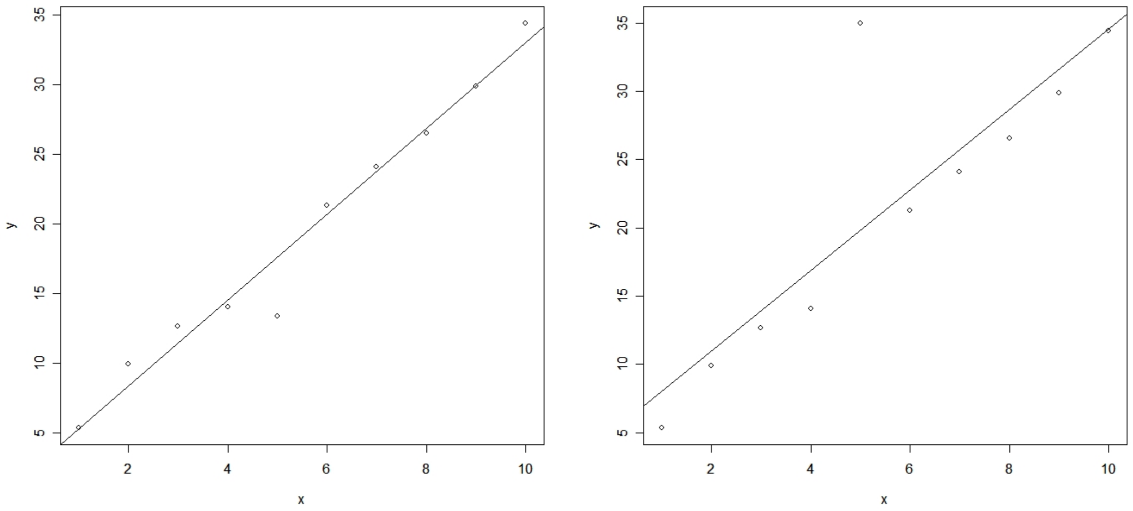
One other thing we can glean from residual plots or residual histograms are whether there are any "outliers" from the model.

An outlier is an extreme observation, different from other observations in the data set. They can usually be identified from the residual plots.

There are formal tests for outliers. For example, one can form "studentized residuals" by taking the set of residuals, subtracting the mean residual (necessarily zero for linear regression models) and dividing each by the residual standard deviation. Then, one can (somewhat arbitrarily) decide that an outlier will be any residual value beyond, say, four standard deviations from 0.

When one finds an outlier, one can do several things:

- Go back and verify in the data set that this value is not simply an error in data entry or
- Delete the outlying observation from the data set and rerun the analysis without that point (or points, if more than one outlier is identified). This allows one to check the influence of any outlier on the estimated coefficients, as they can be thrown off by outliers, as illustrated below.



  Note that both the intercept and slope have substantially changed.

- Realize that there can be unique individuals that do not follow the same patterns as others. For example, individuals with certain diseases, or Olympic athletes may differ in some variables, and so on. Depending on the clinical question, it may or may not make sense to delete these subjects.
- There are some formal tests for outliers, or measures such as Cook's distance, DFBETAS (standardized difference in beta coefficients with and without the outlier), and so on. See any regression textbook for details.

R^2: A global measure of "variance explained"

You may have noticed that the R^2 value is routinely given in the R software outputs, as in almost any statistical software that includes regression. We will now see exactly how to interpret this measure.

Consider first simple linear regression. We are usually interested in whether the independent variable is worth having, so really we are comparing the model

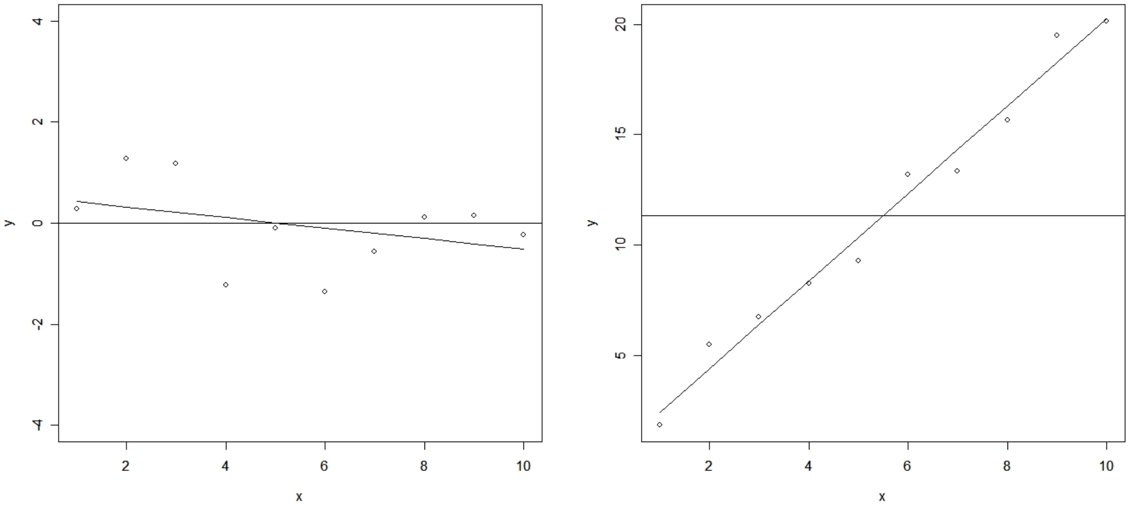
$$Y = \alpha + \beta X$$

to the simpler model

$$Y = \alpha$$

using the intercept only.

Compare the sum of squared residuals in the following two graphs:



In the first case, the independent variable X does not seem to have a large effect, while in the second, the linear model fits well.

We define:

$$R^2 = 1 - \frac{\text{sum of squared residuals from model with } \alpha \text{ and } \beta}{\text{(sum of squared residuals from model with } \alpha \text{ only)}}$$

You often see this notation:

$$R^2 = 1 - \frac{SS(res)}{SS(total)}$$

where $SS(res)$ (often also referred to as SSE or “sum of squares of error”) is defined as the sum of residual distances from model with X, and $SS(total)$ is the same, but from the intercept only model.

By definition of least squares regression, $SS(res) \leq SS(total)$, because if the best regression line was really using α only, then $SS(res) = SS(total)$, and in all other cases, adding β improves $SS(res)$.

So, $0 \leq R^2 \leq 1$. If $SS(res) = SS(total)$, then $R^2 = 0$, and model is not useful. If $SS(res) = 0$, then $R^2 = 1$, and model fits all points perfectly. Almost all models will be between these extremes.

So, $SS(res)$ shows how much closer the points get to the line when β is used, compared to a flat line using α only (which is always $Y = \alpha = \overline{Y}$).

Because of this, we can call R^2 the "proportion of variance explained by adding the variable X".

Essentially the same thing happens when there is more than one independent variable, except residuals are from the model with all X variables for the numerator in the definition of R^2. Thus, R^2 gives the "proportion of variance explained by adding the variables $X_1, X_2, \ldots, X_p$ if there are p independent variables in the model.

Interesting point: If you take the square root of R^2, you get the usual correlation coefficient between X and Y, except possibly for the sign, which is lost in taking the square. This is true regardless of the value of p, the number of parameters in the model, and when $p > 1$, it is called the coefficient of multiple correlation.

How large does R^2 need to be to be considered as "good"? This depends on the context, there is no absolute answer here. For hard to predict Y variables, smaller values may be "good".

Overall, R^2 provides a useful measure of how well a model fits, in terms of (squared) distance from points to the best fitting line. However, as one adds more regression coefficients, $R2$ *never* goes down, even if the additional X variable is not useful. In other words, there is no adjustment for the number of parameters in the model.

Adjusted R^2

In a simple linear regression, where p, the number of independent variables is one, then Adjusted $R^2 = R^2$. As the number of parameters increases, Adjusted $R^2 \leq R^2$, with this definition:

$$R^2 = 1 - \frac{(n-1) \times \text{sum of squared residuals from model with } \alpha \text{ and } \beta}{(n-p) \times \text{sum of squared residuals from model with } \alpha \text{ only}}$$

So, there is some attempt to adjust for the number of parameters.

Example for R^2 and Adjusted R^2

In a previous regression of Heart rates, we saw:

```
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = HR ~ Block + Height + Frequency, data = heart.dat)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8750  -4.3500   0.4125   6.2125   9.8750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   80.425      3.952  20.349 2.65e-15 ***
Block2        -1.925      4.768  -0.404  0.69050
Block3        -4.525      4.800  -0.943  0.35660
Block4        23.125      4.800   4.817 9.24e-05 ***
Block5        17.225      4.768   3.612  0.00163 **
Block6        -5.550      4.752  -1.168  0.25592
Height1       21.750      2.781   7.821 1.18e-07 ***
Frequency1     9.250      3.406   2.716  0.01294 *
Frequency2    24.875      3.406   7.304 3.43e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 7.462 on 21 degrees of freedom
Multiple R-Squared: 0.9035,     Adjusted R-squared: 0.8668
F-statistic: 24.59 on 8 and 21 DF,  p-value: 5.316e-09
```

R reports $R^2 = 0.9035$ and Adjusted $R^2 = 0.8668$, so, in either case, about 90% of the total variance is explained by the three variables used, which is very high. At least by these measures, the model fits well.

Chi-square goodness of fit tests

You may recall that we defined a χ^2 test when discussing testing for proportions as:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Similar tests can be constructed for regression models, where one can create categories based on dividing up the Y axis (i.e., the range of the outcome), and comparing how many observations fall into each category to what is predicted by the model.

While one sometimes sees such tests used, they suffer from all of the usual problems of p-values. For example, since one knows that "all models are wrong", why test the null hypothesis that the model is exactly correct versus the alternative that the model is not perfect? One knows ahead of time that the alternative hypothesis is true, and the test fails to provide any real useful information.

Overall, while one sometimes sees such tests used, they are of questionable utility.

Model validation

One often attempts to "validate" a model either by:

1. Randomly splitting an existing data set into two parts, and using part of the data for "model fitting", and part of the data for "model validation".
2. Using one full data set for "model fitting", and finding a second independent data set for "model validation".

While at first glance intuitively appealing, splitting a data set is of highly questionable utility, for two main reasons:

1. Using only part of data set to develop a model is wasteful of information, and drastically decreases accuracy in estimating any coefficients, changes model selection (smaller models are typically found because of a loss of "power"). In summary, not using all the data leads in general to a poorer model.
2. It is not clear exactly what is being "validated". If the data set is divided randomly into two parts, then, by definition of the randomization process, both sets are equivalent, and so *must* be the same. Any lack of validation must either be from bad luck during randomization, or lack of a large enough sample size.

Overall, it is better to use all the data to develop the best model possible.

Obtaining a new data set improves on the idea of splitting the data set, because it allows for checking of the model in a *different* situation.

If the situation were not different, then the arguments against data splitting would apply to this case as well; one would be better off building a single model using the two assumed very similar situationed data sets.

If the two contexts from which the two data sets arose were different, then, at least, one can check how well the first model predicts observations from the second context. If it does fit, there is some assurance of generalisability of the first model to other contexts. If the model does not fit, however, one cannot tell if the the lack of fit is owing to the different contexts of the two data sets, or true "lack of fit" of the first model.

In practice, these types of validation can proceed by deriving a model and estimating its coefficients in one data set, and then using this model to predict the Y variable from the second data set. One can then check the residuals, and so on.

Final Note

Do not forget, amidst all these statistical ideas, that substantive knowledge and knowledge about a study design can and should play a role in thinking about a model, and how well it suits a given purpose.

If you have good *a priori* reasons to believe a variable should be in a model then simply include it, unless the evidence against it is very strong.

If the main point of a model is prediction, you might not care too much about which independent variables are included, as long as the model "fits well". But if the purpose of your model is to see which variables are important, then much attention needs to be paid to this issue.

Goodness of fit is closely related to model selection, which we will cover in the next lecture.