Regression and Multivariate Data Analysis Summary

Rebecca Sela

February 27, 2006

In regression, one models a *response* variable (y) using *predictor* variables ($x_1, ...x_n$). Such a model can be used for prediction, forecasting, understanding relationships, testing, classifying, and so on.

1 The Linear Regression Model

The linear least squares regression model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i$$

(Note that this is technically a regression hyperplane, not a regression line.) This model is estimated by choosing estimates $\hat{\beta}_0, ..., \hat{\beta}_p$ to minimize sum of squared residuals, $\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_p x_{pi}))^2$. In this model, $\hat{\beta}_j$ is the estimated expected change in the response variable associated with a one-unit change in the j^{th} predictor, holding all the other predictors constant (this is a *partial association*, so the estimates will depend on all the variables in the model). $\hat{\beta}_0$ is the estimated expected value of the response variable when all the predictors equal zero (this may be irrelevant). The fitted values are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_p x_{pi}$. Note that the fitted values may also be written in matrix form as $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$.

We make the following assumptions about the model:

- This straight line relationship is approximately correct.
- $E(\epsilon_i) = 0$ for all i.
- Homoskedasticity: $Var(\epsilon_i^2) = \sigma^2$ does not depend on the observation.
- $Cov(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$.
- $\epsilon_i \sim Normal(0, \sigma^2)$ (or at least approximately so).

Under these assumptions, least squares estimation is optimal. If some of the assumptions are violated, it may be unbiased but not optimal, or entirely wrong.

Consider the identity:

$$\sum(y_i - \overline{y})^2 = \sum(\hat{y}_i - \overline{y})^2 + \sum(y_i - \hat{y}_i)^2$$

where $\overline{y}$ is the mean of the response variable. (This holds only for linear least squares models.) The left-hand side measures the total variability of y_i. The first term on the right-hand side is the variability in y_i explained by the regression model (the explained sum of squares). The second term on the right-hand side is the amount of variability that is not explained by the regression model (the residual sum of squares). We define $R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$. This is the proportion of variability accounted for by the regression. (Closer to one is better, but there is no measure of how good a regression is using R^2.) The adjusted R^2, an unbiased measure of the squared population correlation coefficient, is $R_a^2 = R^2 - \frac{p}{n-p-1}(1 - R^2)$. (This is a small adjustment if n is large.)

We may use hypothesis testing to check the usefulness of the regression. An *overall F-test* tests the null hypothesis that all the coefficients on the predictors equal 0. A *t-test* tests the null hypothesis that an individual coefficient equals zero, given all the other predictors in the model. In the one variable case, the p-values are exactly identical. We may also construct confidence intervals for the regression coefficients: $\hat{\beta}_j \pm t_{\alpha/2}^{n-k-1} se(\hat{\beta}_j)$.

A *linear contrast* is a set of linear equations among the parameters; it is a restriction of the model being estimated. We may use a *partial F-test* to test this restriction, using the residual sum of squares for the full (unrestricted) model and the restricted (subset) model:

$$F = \frac{(RSS_{subset} - RSS_{full})/d}{RSS_{full}/(n-p-1)} = \frac{(R^2_{full} - R^2_{subset})/d}{(1 - R^2_{full})/(n-p-1)}$$

where d is the number of restrictions. Under the null hypothesis, $F \sim F_{d,n-p-1}$. Unless we can reject the null hypothesis of a simple model, we use the simpler model. (Parsimonious models make fewer assumptions about the relationships among all the variables and are often better at forecasting because of it.)

We may understand the practical importance of the regression using prediction intervals; if the prediction intervals are quite narrow, then the regression is useful. A rough prediction interval is given by the predicted value $\pm 2\hat{\sigma}$, where $\hat{\sigma}$ is the standard error of the regression. (The precise standard error of a prediction depends on the distance of the predictors of interest from the mean of the predictors used in estimation.) If we have a large number of observations, the prediction interval is determined almost entirely by the standard deviation of ϵ; for small samples, the uncertainty of the coefficients matters as well. This means there is a non-zero lower bound on the width of the prediction interval (this is in contrast to confidence intervals for the mean value for a given set of predictors, which shrinks to a point).

1.1 Regression Diagnostics

Definition An *outlier* is an observation in which the value of Y is unusual, given the value of X.

Definition The *residual* from a regression is $y_i - \hat{y}_i$; this is the difference between the observed and predicted value. The *standardized residual* is given by

$$e_i^* = \frac{y_i - \hat{y}_i}{se(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where h_{ii} is the (i, i) element from the matrix $H = X(X^T X)^{-1} X^T$.

If the assumptions of the regression model hold, then the standardized residuals should be normally distributed with mean 0 and standard deviation 1. Observations with standardized residuals outside of 2.5 or 3 (especially if the sample size is small) are outliers.

Definition A *leverage point* is an observation in which the values of X are unusual.

Definition The *leverage value* of a point is the effect it has on its own fitted value. This is h_{ii} (since $\hat{y}_i = \sum_{j=1}^n h_{ij} y_j$, h_{ii} is the effect of y_i on $\hat{y}_i$).

Note that $0 < h_{ii} < 1$ and $\sum_{j=1}^n h_{ij} = p + 1$. Thus, we have the guideline that a point may be a leverage point if $h_{ii} > 2.5\frac{p+1}{n}$. Plots of the leverage values for all the points can help identify which points have the most leverage, even if they do not meet this threshold.

Leverage points tend to draw the regression line toward them, making them overly influential. Because of this, a small change in the response for the leverage point causes large changes in the regression line itself. A leverage point that has the same relationship to X as all the other points will tend to decrease standard errors and increase measures of fit. However, it is hard to know if the relationship is *really* the same.

Definition *Cook's distance* is given by $D_i = \frac{(e_i^*)^2 h_{ii}}{(p+1)(1-h_{ii})}$. This is the amount that the coefficients change if the i^{th} point is excluded from the estimation.

Cook's distance increases with both the size of the residual and the amount of leverage of a point. In general, if $D_i > 1$ or if one point has a higher Cook's distance than all the others, that point should be more closely examined.

Definition The *deleted residual* is the residual an observation would have if the regression were based on the other observations and its value were then predicted.

In general, the larger the sample size, the smaller the effect of individual observations. Identifying outliers and the reasons they are unusual can be helpful in understanding the underlying problem.

Definition When there are multiple outliers or leverage points, so that diagnostics based on a single point (such as Cook's distance) cannot catch them, we call this *masking*.

Definition When outliers lead to the "good" points having the highest Cook's distances, then we call this effect *swamping*.

To deal with such problems, we may either use robust regression methods (which are less susceptible to outliers) or use diagnostics that test groups of points. We may also try to fit the regression with a group of points that are known not to have leverage points or outliers (perhaps identified using clustering methods) and then add in those observations that have small predictive residuals based on the regression on the "good" points.

There are also graphical methods of checking the regression assumptions:

- Plot Y versus each predictor. Make sure there is no non-linearity or heteroskedasticity.
- Plot the residuals versus the fitted values, to make sure there is no pattern (isolated points – either horizontally or vertically – can be a problem; so can a width that changes with the fitted values).
- If the observations have a natural order (such as time), plot the residuals versus the order and ensure that there are no patterns.
- Look at a normal probability plot (a plot of the ordered residuals versus the quantiles of the normal distribution). Curves in this plot suggest non-normality, and isolated points are outliers.

Multicollinearity occurs when some of the predictors are highly correlated with each other. In this case, it is hard to isolate the effect of one variable relative to another, and the regression estimates may be unstable. In addition, the regression may be significant overall, even though no individual coefficient is statistically significant. We quantify multicollinearity using the *variance inflation factor* for each variable, defined by:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2 from the regression of the j^{th} predictor on all the other predictors. High values of the variance inflation factor signal multicollinearity. In general, if $VIF_j < \max\{10, \frac{1}{1-R^2_{model}}\}$ for all j, then this should not be a problem.

1.2 Hierarchical Models with Dummy Variables

Suppose we have a response variable, Y, an observed variable, X, and a dummy variable, D. The dummy variable describes a pair of subgroups in the data. Then, we have three possible models which are nested:

- Pooled Model: $Y = \beta_0 + \beta_1 X + \epsilon$ in which the model does not change across the two subgroups.
- Constant Shift Model: $Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon$. In this case, the intercept differs across the two groups, but the slope is the same in the two subgroups. (We test this model versus the pooled model using a null hypothesis that $\beta_2 = 0$.)
- Full Model: $Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX + \epsilon$. In this model, both the slope and the intercept may differ across the two groups. This is equivalent to estimating two regressions (constrained to have the same error variance) and allows us to test the hypotheses that either the pooled or constant shift models are sufficient.

The term XD in the full model is called an *interaction effect*.

1.3 Nonlinearities and Transformations

Note that a linear model is linear in the parameters. We may use non-linear functions of Y or X in a linear model, though it will affect the interpretation of the coefficients.

Some models are inherently non-linear, and there is no way to fit them using linear regression (other methods, such as non-linear least squares or maximum likelihood, may be more useful here).

Definition A model is *linearizable* if it is non-linear but can be transformed into a linear model.

Definition The *log/log model* is defined by $Y = \alpha X^{\beta}$, so that $\log(Y) = \log(\alpha) + \beta \log(X)$, which is a linear regression model. (Note that the base of the logarithm does not matter, as long as it is consistent throughout the analysis.)

Log-log models are commonly used with data involving money, several orders of magnitude, or long right tails. In this model, β is the *elasticity* of Y with respect to X; that is, β is the estimated percentage change in Y associated with a 1% change in X.

For any model with $\log(Y)$ on the left-hand-side, R^2 measures the variation in $\log(Y)$ explained by the right-hand-side (not the variation in Y itself!). Any prediction intervals for $\log(Y)$ may be converted into prediction intervals for Y by exponentiating the endpoints of the interval.

Definition The *semi-log model* or the *additive/multiplicative model* is given by $Y = \alpha \gamma^X$, so that $\log(Y) = \log(\alpha) + X \log \gamma = \beta_0 + \beta_1 X$. Then, β_1 is a *semielasticity*, which measures the multiplicative change in Y associated with an additive change in X.

Multiplicative/additive models, with $Y = \beta_0 + \beta_1 \log(X) + \epsilon$, are also useful in some cases.

1.4 Model Selection

Model selection is the problem of choosing which of a set of variables belong in a linear regression. The hope is to choose a parsimonious yet accurate model and to avoid multicollinearity.

First, we may use a *best subsets regression*, in which the computer efficiently fits all possible models and displays the best model(s) for each number of variables (based on R^2, which is equivalent to most other methods when the number of variables is fixed). Note that the best models of different sizes need not be nested, which means we may not use F- or t-tests to choose among them. Instead, we have the following measures which may help us trade off:

- R^2: Note that adding more variables never decreases R^2. However, we may be interested in the number of variables where R^2 "levels off," since the additional variables are not adding much explanatory power.
- Adjusted R^2: This statistic is given by $R_a^2 = R^2 - \frac{p}{n-p-1}(1-R^2)$ and therefore penalizes models with more variables. We may then choose the model that maximizes R_a^2. (This tends to choose larger models.)
- Standard error of the estimate (s): The standard error of the estimate is proportional to the width of the confidence intervals, so we may wish to choose the number of variables where s levels off or starts increasing again.
- Mallow's C_p: This is defined as $C_p = \frac{ResidualSS}{s_0^2} - n + 2p + 2$, where s_0^2 is the residual mean square from the model with all the predictors. Note that $C_p \approx p+1$ if the residual sum of squares from the model is close to the residual sum of squares from the model with all the variables, so this is a more parsimonious model with about the same residuals. That is, we choose the smallest model where $C_p < p+1$. We may also choose the model that minimizes C_p.
- Akaike Information Criterion (AIC): We define $AIC = n\ln(\frac{ResidualSS}{n}) + n + 2p + 4$ and choose the model that minimizes this value.
- Corrected Akaike Information (AIC_C): This is defined as $AIC_C = AIC + \frac{2(p+2)(p+3)}{n-p-3}$, and corrects the bias in AIC; we choose the model that minimizes this number.

C_p, AIC, and AIC_C are *efficient* model selection methods; they will find the best model (as $n \to \infty$) under the assumption that the true model is never an option. (In contrast, BIC is *consistent* and will find the true model given that the true model is an option.)

Once a few models are chosen by these criteria, the assumptions can be checked for those models and a single model can be chosen based on both the criteria above and other regression diagnostics.

However, by running so many possible regressions, we risk finding relationships by chance. To avoid this, we may reserve some data to validate the model

based on the "training data" (this also allows a better estimate of s^2). We may also just adjust the standard error of the estimate to account for the number of variables that were considered: $s^* = \sqrt{\frac{ResidualSS}{n-\max(p)-1}}$. Ideally, we should correct the t- and F-statistics as well. We may also randomly perturb the data and see if a similar model is chosen (the method *bagging* does this repeatedly and then averages predictions over the set of models that were chosen). Finally, it is worth thinking about which variables should be used instead of just throwing in lots of variables.

1.5 Partial Correlation

The coefficient on a variable, X_1, in a model measures the association between Y and X_1 controlling for all the other variables in the model. We may see the relevant relationship in this way:

1. Regress Y on all the other variables, $X_2, ..., X_p$, and save the residuals, e_Y.
2. Regress X_1 on all the other variables, $X_2, ..., X_p$, and save the residuals, e_1.

The *partial correlation* is the correlation between e_1 and e_Y. The regression of e_Y on e_1 will have the same coefficient as the coefficient on X_1 in the original regression. A scatterplot of e_Y versus e_1 may also show patterns of interest.

1.6 Time Series Data

Suppose that observations are ordered in time. Then, they are likely to to violate the assumption that $Cov(\epsilon_s, \epsilon_t) = 0$, which will make the estimates inefficient and the measures of fit misleading (and overly optimistic, when both ϵ and X have positive autocorrelation).

A plot of residuals versus the order of observations will tend to show "cyclical" patterns (where the residuals tend to stay on the same side of zero for a while).

One formal test for autocorrelation is the *Durbin-Watson Test*. In this case, we test $H_0 : \epsilon_i \sim Normal(0, \sigma^2), Cov(\epsilon_i, \epsilon_j) = 0$ versus $H_A : \epsilon_i = \rho\epsilon_{i-1} + z_i, z_i \sim Normal(0, \tau^2)$. The Durbin-Watson test statistic is $DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$, where e_i are the residuals. $E(DW) = 2$ under the null hypothesis, and DW is less than two if there is positive autocorrelation. For large samples ($n > 100$), $\sqrt{n}(\frac{DW}{2} - 1) \sim Normal(0, 1)$, approximately. For smaller samples, we may compare $4 - DW$ to a table. Note that this test may be indeterminate (because it technically should depend on the values of X), and one might want to correct for autocorrelation in the indeterminate region. This test is the most powerful if H_0 and H_A are correct, but may fail if the assumptions are violated.

We may also plot the *autocorrelation function*. This allows us to see whether any of the autocorrelations are non-zero (as well as see whether the assumptions

of the Durbin-Watson test are approximately met; in that case, the autocorrelations should decay exponentially). Note that the hypothesis tests about the ACF assume approximate normality.

The *runs test* is a non-parametric test of the null hypothesis that there is no autocorrelation in the errors. Suppose we observe n_+ positive residuals, n_- negative residuals, and u runs of consecutive residuals with the same sign. Under the null hypothesis,

$$\frac{1}{\sigma}(|u-\mu|-\frac{1}{2}) \sim Normal(0,1)$$

where $\mu = \frac{2n_+n_-}{n_++n_-} - 1 \approx \frac{n}{2} + 1$ and $\sigma^2 = \frac{1}{n^2(n-1)}(2n_+n_-(2n_+n_- - n)) \approx \frac{n^2-2n}{4(n-1)}$. This test is always valid, but is less powerful.

To deal with a time series, we should consider *detrending*, by removing either a linear trend (by adding time or a function of time as a predictor) or some underlying change, like inflation or population growth. If the ACF plot shows a spike at a frequency suggesting seasonality (ρ_4 for quarterly data, ρ_{12} for monthly data), we may *deseasonalize* by adding dummy variables for all but one period to control for seasonality.

We may also *lag* either the independent or the dependent variables (this is only helpful for forecasting if the previous value would be known in time for the forecast!). Lagging the predictors generally doesn't fix autocorrelation, but it allows for effects on the response to take some time (which may make sense depending on the context). Lagging the response is often helpful in fixing autocorrelation in the residuals. (However, including lagged response variables makes the Durbin-Watson test invalid.)

We may also *difference* either the predictors or the residuals. Note that differencing the predictor is context dependent and is really a special case of the model that includes both x_t and x_{t-1} as predictors (where their coefficients are equal with opposite signs). Differencing the response is equivalent to having a lagged variable as a predictor with a coefficient of 1, except that the left-hand-side variable is different (which will affect R^2 and test statistics).

Outliers are harder to omit in time series, especially if there are lagged values in the model. Instead, we wish to *impute* a new value. The imputed value can come from:

- Linear Interpolation: $y_t^* = \frac{1}{2}(y_{t-1} + y_{t+1})$ (this does not consider the predictor variables)
- We may estimate the model without using y_t as an observation or a predictor and use the fitted value, $\hat{y}_t$.
- Multiple Imputation: Impute multiple values of y_t^* and compare the analyses across the different values.

We may also use a model-based method that includes an indicator that is 1 during the outlier period and 0 otherwise (this can also help account for groups of consecutive outliers from the same cause). The coefficient on the dummy

variable estimates the effect at that period, holding all else equal. (If there are multiple groups of outliers, there should be different indicators for each.)

We may also deal with autocorrelation structure in the noise. Suppose we have $y_t = \beta_0+\beta_1 x_{it}+...+\beta_k x_{kt}+\epsilon_t$ with $\epsilon_t = \rho\epsilon_{t-1}+z_t$ (and the z_t independent). For this, we use the *Cochrane-Orcutt Procedure*:

- If we let $y_t^* = y_t - \rho y_{t-1}$ and $x_{jt}^* = x_{jt} - \rho x_{j,t-1}$, then we note that $y_t^* = (1-\rho)\beta_0 + \beta_1 x_{1t}^* + ... + \beta_k x_{kt}^* + z_t$ and OLS is optimal in this case.
- Since we do not know ρ, we estimate $\hat{\rho}$ from the first autocorrelation of the errors in the original regression.
- Transform $y_t^* = y_t - \hat{\rho} y_{t-1}$ and $x_{jt}^* = x_{jt} - \hat{\rho} x_{j,t-1}$.
- Regress y_t^* on $x_{1t}^*, ..., x_{kt}^*$. and do all the diagnostics on the residuals from this regression. Note that the existence of autocorrelation in these residuals suggests that modeling the errors as AR(1) was inadequate.
- To relate the transformed regression to the original regression, the slope coefficients are identical, $\hat{\beta}_o = \frac{\hat{\beta}_0^*}{1-\hat{\rho}}$, and $\hat{\sigma} = \frac{\hat{\sigma}^*}{\sqrt{1-\hat{\rho}^2}}$. (Note that the prediction intervals are wider in the original model.)

1.7 Heteroskedasticity

1.7.1 Heteroskedasticity Related to the Response Variable

Suppose that we have heteroskedasticity that is related to the level of Y (not to the explanatory variables). Then, there may be a *variance-stabilizing transformation*, h, such that $h(y)$ has a constant variance. Suppose that $\sigma_y^2 \propto f(\mu_y)^2$. Then, choose a function, h such that $h'(\mu_y) \propto \frac{1}{f(\mu_y)}$. Then, by a Taylor series expansion:

$$Var(h(y)) \approx (h'(\mu_y))^2 Var(y) \propto (h'(\mu_y))^2 f(\mu_y)^2$$

which is approximately constant. In particular, suppose $\sigma_y = \mu_y^k$. Then $h(y) = y^{1-k}$. (If $k = 1$, then $h(y) = \ln(y)$.) The transformation may be context dependent. For example, common transformations include:

- Poisson (count) data: If $Y \sim Poisson(\lambda)$, then $E(Y) = Var(Y)$ and $\sigma_y \propto \mu_y^{1/2}$. Then, $h(y) = \sqrt{y}$ and we should analyze the square roots of the data instead.
- Gamma (fixed scale) data: In this case, $h(y) = \ln(y)$. Note that this is related to the previous situation in which we took logs; the distribution has a long right tail and a fixed scale (that is, $\sigma/\mu = \frac{1}{\sqrt{\alpha}}$ is constant).
- Exponential (waiting times): In this case, $\frac{1}{y}$ is usually used (which means that we are modeling something like the number of events per time instead of the time to the next event).

These transformations allow us to use OLS, but the meanings of the coefficients change. It may be better to use other kinds of regression (general linear models) instead.

1.7.2 Heteroskedasticity Related to the Predictors

Suppose $y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i$ and $Var(\epsilon_i) = \sigma_i^2$. Then, OLS of $\frac{y_i}{\sigma_i}$ on $\frac{1}{\sigma_i}, \frac{x_{1i}}{\sigma_i}, ..., \frac{x_{ki}}{\sigma_i}$ is optimal because the variance of the errors is now constant (this is equivalent to *Generalized Least Squares*, which weights observations by $\frac{1}{\sigma_i}$). These estimates are more efficient and lead to more accurate prediction interval widths. However, the σ_i and therefore the weights are usually unknown.

Levene's Test for Heteroskedasticity:

1. Use OLS to estimate the absolute standardized residuals.
2. Regress the absolute residuals on the predictor variables that are possibly related to heteroskedasticity (these variables need not be in the regression for the levels).
3. Use an overall F-test to test whether the variables are significantly related to the variability of the residuals.
4. Individual t-tests may also help identify which variables are most important.

Alternatively, we may plot the absolute residuals versus a variable and find a non-parametric curve for the data (using a Lowess curve); this may suggest a suggest a functional form for Levene's test.

Correcting for heteroskedasticity related to continuous variables:

- Assume that $\sigma_i^2 = \sigma^2 \exp(\sum_j \lambda_j z_j)$ where the z_j are the variables affecting heteroskedasticity.
- Estimate the coefficients, λ_j, using the regression of $\ln(\hat{\epsilon}_i^2)$ on z_j.
- The estimated weights are $\frac{1}{\exp(\sum_j \hat{\lambda}_j z_j)}$.

After this correction, the standardized residual plots should not have heteroskedasticity, in either plots or Levene's test (the standardized residuals should correct for heteroskedasticity). In GLS, the coefficients have the same meaning but the values might change. The standard error no longer has any meaning (since each observation has its own standard error), nor does R^2. The influential observations may change, since their weights affect their influence.

To predict with weighted least squares, we use the same method to find point estimates. However, the prediction interval width will vary with the weight, We estimate the weight using the previously estimated relationship, $\frac{1}{\exp(\sum_j \hat{\lambda}_j z_j)}$, with the z values from the new observation. The correct standard error for this predicted value is $se^* = \sqrt{(se_{fit})^2 + (ResidualMS)/Weight}$, which can be

used to create a prediction interval. (Note that se_{fit} is given correctly in the output because it does not depend on the weight of an individual observation.)

We may also have non-constant variance within subgroups only. We may calculate the residuals from OLS and calculate summary statistics for each subgroup. Then, the weight for each observation in group j is $\frac{1}{\hat{\sigma}_j^2}$. This is a non-parametric method, but it works only with discrete variables.

2 ANOVA

2.1 One-Way ANOVA

Suppose we observe a variable across K different groups. Then we model $y_{ij} = \mu+\alpha_i+\epsilon_{ij}$, where μ is the overall mean (assuming the groups have equal weights), α_i is the effect of being in the i^{th} group (so that the group mean is $\mu + \alpha_i$), and $\epsilon_{ij} \sim Normal(0, \sigma^2)$ is the error for the j^{th} individual in the i^{th} group. To identify the model, we impose the restriction that $\sum_{i=1}^{K} \alpha_i = 0$; if there is no group effect, then all of the α_i are 0. We may also write this as a regression model with $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$ where x_i is an indicator variable for group i. Note that we must either (1) omit one indicator (corresponding to the reference group) or (2) use *effect codings* where we instead set $x_i = -1$ for the omitted group. With just indicator variables, the mean for the omitted group is β_0 while the mean for the i^{th} group is $\beta_0 + \beta_i$. With effect codings, the mean for group $i \neq K$ is $\beta_0 + \beta_i$ and the mean for group K is $\beta_0 + \sum_{i=1}^{K-1} \beta_i$; that is, β_0 is the overall level and $\beta_1, ..., \beta_{K-1}, -\sum_{i=1}^{K-1} \beta_i$ are the group effects. Note that the two codings are equivalent, but the interpretation of the coefficients differs.

This method does not impose any relationship between group order and group mean. However, it does impose a constant variance.

We may wish to test whether specific pairs of groups are different; that is, we want to test null hypotheses of that form $\alpha_i = \alpha_j, j \neq i$. With so many possible pairs, we have the *multiple comparisons problem*. Instead, we do simultaneous inference to ensure that the overall error rate is correct. The exact method of controlling the overall error rate is called *Tukey's T*. An approximate method is the *Bonferroni method* which tests at an overall confidence level of $1 - \alpha$ by considering only p-values above α/k significant. If the subgroups are ordered, the we may also wish to test a functional relationship between the subgroup order and the mean; this leads to a regression of y on $f(i)$.

We may also wish to check for non-constant variance by subgroup.

2.2 Two-Way ANOVA

Suppose we have two (or more) categorical predictors. Then the two-way ANOVA model is given by:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

for groups i, j and observation k in the i, j group. We assume that $\epsilon_{ijk} \sim Normal(0, \sigma^2)$. Then, the observed values depend on the overall level, μ, the main effects of the two subgroups ("rows" and "columns") and the interaction effect, $(\alpha\beta)_{ij}$ (this is notation, not actual multiplication). To identify the parameters, we require that:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

The *interaction effect* is the change in the row effect due to the column (or vice versa). An *interaction plot* draws a line for each row with points for each column; if there are no interactions, then the lines in the interaction plot are parallel. We always include the main row and column effects if we include the interaction effects. If the interactions are insignificant, then we may omit the interaction effects and test whether the row and column effects are significant.

To calculate the regression, we may use effect codings for the row and column effects and then calculate the interaction as the product of row and column effects (this is $(I-1)(J-1)$ different pairwise products). In this setup, the test for an interaction effect is a partial F-test that compares the model with and without the interaction effects.

For this model to be estimable, no cell may be empty (that is, we must have at least one observation from every pair (i, j)). If there are empty cells, we may omit rows or columns with missing cells, combine rows or columns so that no cell is empty, or omit interaction effects; otherwise, we have perfect multicollinearity.

It is preferable for all the cells to have approximately equal numbers of observations, but this may not be possible. If n_{ij} is constant for all i, j, then we have a *balanced design*. In this case, the effects are orthogonal and the p-values for row and column effects will change very little when the interaction term is removed. Also, all the points have equal leverage. In a balanced design with no interactions, the fitted values are $\overline{y_{i\cdot}} + \overline{y_{\cdot j}} - \overline{y_{\cdot\cdot}}$. If $n_{ij} = 1$ then a model with interactions will fit the data perfectly.

There may be non-constant variance in the different subgroups, which suggests using weighted least squares. If there is non-constant variance by the two groups (with no interaction), then the weights are given by:

$$w_{ij} = \frac{1}{stdev_i^2} \frac{1}{stdev_j^2}$$

Alternatively, Levene's test may be implemented using ANOVA. If the interaction is significant, then the weights are given by $\frac{1}{stdev_{ij}^2}$ instead.

2.3 Analysis of Covariance

Suppose we have a model with *covariates* (numerical variables), $X_1, ..., X_p$ and *factors* (subgroups), $i = 1, ..., I; j = 1, ...J$. Then, we have an ANCOVA model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_1 x_{1ijk} + ... + \gamma_p x_{pijk} + \epsilon_{ijk}$$

This is a constant shift model, in which each subgroup has a different intercept, so that given certain values for the covariates, the difference between being in groups i and i' depends only on the α terms (and their interactions). This allows for multiple comparisons tests to be used.

The point prediction in this case is $Mean_i + \hat{\beta}_1(x_1 - \overline{x}_1) + ... + \hat{\beta}_p(x_p - \overline{x}_p)$ (the means of the covariates could be combined into a constant term, depending on the output). A rough half-width of the prediction interval is $2\sqrt{\frac{ResidualMS}{Weight_i}}$ (where $Weight_i = 1$ if we are not using WLS).

We may also allow the slopes to differ by subgroup by including interaction terms between the subgroups and the covariates. To calculate the slopes from variables based on effect codings, we add the slope of the i^{th} group to the overall slope (except for the last group, where we subtract the sum of all the other slopes from the overall slope, just as we do for groups means above).

3 Generalized Linear Models

A more general model of regression has three parts: a random component, the expected relationship of y to X, and a link function from the random part to the predictor. In a linear regression, the expected relationship is $\mu_i = \beta_0 + \beta_1 x_i$, the random component is $y_i \sim Normal(\mu_i, \sigma^2)$, and the link function is μ. In a generalized linear model, we assume that y has a distribution in the exponential family, given by:

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

We define $\mu = E(y) = b'(\theta)$ and note that $Var(y) = a(\phi)b''(\theta)$. We define the *systematic component* by:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

The *link function* is given by $g(\mu) = \eta$. The *canonical link function* is $\theta_i = \eta_i$ (this often makes estimation and hypothesis testing easier).

Let $r_i = (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$ and $W = diag[(\frac{\partial \mu_i}{\partial \eta_i})^2 / Var(y_i)]$. Then, the score equations for maximum likelihood estimation are $0 = X'Wr$, and we may estimate $\hat{\beta} = (X'WX)^{-1}X'Wz$ where $z = X\beta + r$. Note that W and r depend on the parameters, so that estimation is done through iteratively reweighted least squares.

3.1 Goodness of Fit Tests

Definition A *goodness of fit* test is a test of the null hypothesis that the model fits the data. (It is not a test of the strength of the relationship.)

Definition The *saturated model* is the model in which we estimate $\hat{\mu}_i = y_i$; this model has one parameter for each observation. The *deviance* is the likelihood

ratio test statistic of the saturated model versus the model of interest:

$$D(y, \hat{\mu}^M) = 1 \sum_{i=1}^{N} \frac{y_i(\hat{\theta}_i^S - \hat{\theta}_i^M) - b(\hat{\theta}_i^S) + b(\hat{\theta}_i^M)}{a_i(\phi)} = \sum_{i=1}^{M} d_i$$

If the model is a good fit, then D χ^2_{n-p-1}.

A second goodness of fit test is given by the Pearson statistic, $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{V}(y_i)}$. If the model is a good fit, X^2 also has a χ^2_{n-p-1} distribution.

We may also use the deviance to compute a "pseudo-R^2":

$$R^2_D = 1 - \frac{D(y, \hat{\mu}^M)}{D(y, \hat{\mu}^0)}$$

where $\hat{\mu}^0$ are the estimates from the model with only an intercept.

3.2 Model Checking

Many of the statistics used for model checking can be used with general linear models.

The leverage values are the diagonal entries of the matrix $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$.

There are two possible forms of residuals:

- Pearson Residuals: $r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}}$
- Deviance Residuals: $r_i^D = sgn(y_i - \hat{\mu}_i)\sqrt{d_i}$

Note that the squared residuals add up to their respective goodness of fit measures. To standardize the residuals, to $\tilde{r}_i^P$ and $\tilde{r}_i^D$, we divide by $\sqrt{\hat{\phi}(1 - h_{ii})}$.

We define Cook's distance (as before) by:

$$CD = \frac{\tilde{r}_i^2 h_{ii}}{(p+1)(1 - h_{ii})}$$

where either type of standardized residuals may be used.

As before, plots of residuals or absolute residuals versus the predictors may be of interest. In addition, plots of the residuals versus $\hat{\eta}$ may be helpful.

4 Logistic Regression

Logistic regression (also called *binomial regression*), is a general linear model with:

$$\begin{aligned} y_i &\sim Binomial(n_i, p_i) \\ \ln(\frac{p_i}{1 - p_i}) &= \beta_0 + \beta_1 x_i \end{aligned}$$

The *odds* are $\frac{p}{1-p}$; there is a one-to-one correspondence between probabilities and odds. Note that logistic regression is a semi-log model of the odds. In this model, we have:

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

According to this model, increasing the j^{th} predictor by one unit multiplies the odds ratio by $e^{\hat{\beta}_j}$. If the confidence interval for $e^{\hat{\beta}_j}$ contains one, then the variable does not have a significant effect.

To plot the data, we may look at boxplots of the continuous predictors versus by outcome (this just reverses the usual axes). The outliers in the boxplots may be interesting points. Note that non-constant variance across the boxplots is no longer relevant to heteroskedasticity, but they may suggest variables for which the logarithm will be a better predictor.

In a logistic regression with categorical predictors, we may consider the data in a contingency table with one dimension for each predictor and one dimension for the response. (This also provides a better way to display the data, and may show whether there is an interaction.)

In testing, we use Wald tests (a normal approximation; exact methods exist but are computationally intensive) for each individual variable. Furthermore, the test for the overall significance of the regression (G) is a likelihood ratio test, which may give slightly different results from the Wald test (in the case where the logistic regression has only one variable).

In logistic regression, useful diagnostics for the model include:

- *Delta Beta*: This is the logistic regression analog of Cook's Distance, which measures the effect of a single observation on the output. (In terms of scale, it is more like the square root of Cook's distance.)
- *Leverage values (H_i)*
- *Estimated Probabilities*: These are the "fitted values", the estimated probability of the event occurring for each observation.
- *Standardized Pearson Residuals*: These are proportional to the observed number of successes less the expected number. (Residuals are positive when there is a success and negative when there is a failure.)
- $\Delta\chi^2$*, the squared Pearson residuals*: A plot of these versus the fitted values can show outliers (those values above 6 or 7). Because the responses are discrete, there will always be an x-shaped pattern in the plot of $\Delta\chi^2$ versus the estimated probabilities, even if there is no structure in the residuals.

To quantify the usefulness of the variables in predicting outcomes, we use *measures of association*. The most common counts the number of concordant, discordant, and tied pairs. This runs through all possible pairs of one success and one failure. If the estimated probability (of success) is greater for the success, the pair is *concordant*; if the estimated probability (of success) is greater for

the failure, then the pair is *discordant*; otherwise, it is a tie. A model with more concordance is generally preferable (this can help compare two models with the same number of variables). The number of concordant and discordant pairs is summarized by various statistics, including Somer's D. Other measures of association exist as well.

To test the usefulness of a model, we may create a table where the rows are the actual successes and failures and the columns are the predictions (where we predict "success" if the estimated probability is above some fixed number, like 0.5 or another number if one kind of misclassification is worse than the other). Then, the percent of correct predictions is $\frac{n_{00}+n_{11}}{n}$. We may compare this number against:

- $C_{max} = \max(\frac{n_{0\cdot}}{n}, \frac{n_{1\cdot}}{n})$: This is the probability we would have correct if we classified everything into the larger group. (This is harder to beat if most of the outcomes are of the same type.)
- $C_{pro} = 1.25(\frac{n_{0\cdot}n_{\cdot 0}}{n^2})(\frac{n_{1\cdot}n_{\cdot 1}}{n^2})$: If the predictors are useless, then the predicted and actual results would be independent, and we would expect $n_{00} = \frac{n_{0\cdot}n_{\cdot 0}}{n}$ and $n_{11} = \frac{n_{1\cdot}n_{\cdot 1}}{n}$. We inflate this by a factor of 1.25, since the regression used the same data for estimation and prediction.

Collinearity can occur in a logistic regression, but it is not defined in the same way.

Note that the model is estimated iteratively, and might not converge. If the model doesn't converge, then the results are not reliable. Sometimes, one might just need more iterations. If it still doesn't converge, the data might fit the model perfectly. In this case, all the estimated probabilities are zero or one. This is called *separation*, because the successes and failures can be perfectly separated by the variables. To fix this, some variables should be omitted from the model.

Another option is probit analysis, where the link function is the CDF of a standard normal. However, the logit is canonical link function and the estimates have better small sample properties and are fully efficient.

4.1 Goodness of fit tests

The binomial distribution depends on only one parameter, p_i, instead of two, μ, σ^2. This means we do not need to estimate the variance separately, which allows for goodness of fit tests.

Lack of fit might be caused by outliers, the wrong functional form, or the wrong link function (perhaps there is *unmodeled heterogeneity* which leads to overdispersion, with a variance bigger than $np(1-p)$; this can be caused by trials that are not independent, omitted variables, and clustering). If the model does not fit, it may be able to be modified (with a Beta-Binomial, for example) or quasi-likelihood or semiparametric methods may be more useful.

If we observe multiple trials for each i (preferably, $n_i > 5$ for all i), then we have two asymptotically equivalent goodness of fit tests:

- Pearson Goodness of Fit Test: $X^2 = \sum_j \frac{(f_j - n_j\hat{p}_j)^2}{n_j\hat{p}_j(1-\hat{p}_j)}$, where f_j is the observed number of successes in n_j trials and $\hat{p}_j$ is the estimated probability.
- Deviance Test: $G^2 = 2\sum_j f_j \ln(\frac{f_j}{n_j\hat{p}_j}) - (n_j - f_j)\ln(\frac{n_j f_j}{n_j(1-\hat{p}_j)})$

Under the null hypothesis of a correct model, both test statistics have a χ^2_{N-p-1} distribution asymptotically (they will be zero if $\hat{p}_j = \frac{f_j}{n_j}$), where N is the number of groups of observations.

If the n_i are small, then we use the *Hosmer-Lemeshow* test. This test ranks the data by $\hat{p}$, splits the data into roughly equal-sized groups by rank, calculates the expected number of successes in each group, and runs a goodness of fit test based on these numbers.

If the goodness of fit test rejects the model, other predictors or interactions may be necessary (ignoring covariates can cause big problems, like Simpson's paradox). For large n_i, we may graph the *empirical logits*, $\ln(\frac{f_j/n_j}{1-f_j/n_j})$, versus the predictor to check for non-linearity. (If this cannot be calculated because of a zero, we made add a fake success and a fake failure to make calculation possible.)

4.2 Model Selection

To choose a model, we have a variety of possible measures:

- G measures the overall strength of the regression (higher values are preferable). However, G always increases with additional predictors.
- Somer's D: The model where this measure begins to level off is best.
- A model with a high p-value of the Hosmer-Lemeshow test is preferable, because the model fits better.
- Minimize $AIC_C = G^2 + 2\nu(\frac{n}{n-\nu-1})$, where G^2 is the deviance and ν is the number of parameters.

Instead of using best subsets for a logistic regression, we may approximate the process by using best subsets with a linear regression of the 0-1 values on the predictors. (This tends to work better if most of the estimated probabilities are away from 0 and 1.)

If we drop variables based on the Wald statistics (since they are only approximate), one should check that the new model still fits (using the Hosmer-Lemeshow test or the Pearson and Deviance tests).

4.3 Adjusting the odds for data collection

Definition A *prospective (cohort) study* is a study in which units are sampled at random and then followed to see the outcome. (This can be done in real time or just by choosing a random sample from a valid frame in the past.) Note that this sample is not based on the outcome of interest.

Definition A *retrospective (case-control) study* samples a fixed number from each outcome of interest in order to measure other variables. This sample is based on the outcome of interest.

Retrospective studies ensure a more equal representation of the two outcomes, but estimates for future individuals need to be adjusted for the sampling method. (Logistic regression allows for comparison between these two types of studies.)

¿From a prospective study, we may calculate the odds of observing the outcome given the predictor, $\pi_{O|P}$. From a retrospective study, we calculate the odds of the predictor given the outcome, $\pi_{P|O}$. The logistic regression uses the *odds ratio*, which we may calculate from a prospective or a retrospective study:

$$\frac{\pi_{O|P}/\pi_{\sim O|P}}{\pi_{O|\sim P}/\pi_{\sim O|\sim P}} \frac{\pi_{P|O}/\pi_{\sim P|O}}{\pi_{P|\sim O}/\pi_{\sim P|\sim O}}$$

Logistic regression is based on the odds ratio, which means that the coefficients from the retrospective study and a prospective study are comparable. However, the intercepts will differ because of the way the data was collected.

In a prospective study, we estimate the probabilities for new observations by:

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}$$

In a retrospective study, we must know the unconditional proportions, $\pi_O, \pi_{\sim O}$, of each outcome (from some other data source). Then we use the ratio between the number of each outcome observed in the study ($n_O, n_{\sim O}$), and these probabilities to adjust the intercept:

$$\tilde{\beta}_0 = \hat{\beta}_0 + \ln(\frac{\pi_O/n_O}{\pi_{\sim O}/n_{\sim O}})$$

The slope stays the same, and the probabilities are calculated as above.

4.4 Multinomial Logistic Regression

We use multinomial logistic regression when the target variables have more than two outcomes. These outcomes may be *nominal*, where there is no ordering to the outcomes, or *ordinal*, where there is a natural order to the outcomes (but OLS is not appropriate, since the outcomes must be integers).

4.4.1 Nominal Multinomial Logistic Regression

In the nominal case, we have the model for outcomes $j = 1, ..., K$:

$$P(j|X) = \frac{\exp(\beta_{0j} + \beta_{1j}x_1 + ... + \beta_{pj}x_p)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{0l} + \beta_{1l}x_1 + ... + \beta_{pl}x_p)}$$

or, equivalently,

$$\ln(\frac{p_j}{p_K}) = \beta_{0j} + \beta_{1j}x_1 + ... + \beta_{pj}x_j$$

We call K the *control* or *reference group*. This is like fitting $K-1$ logistic regressions, for each of $\ln(\frac{p_j}{p_K})$. We fit them simultaneously so that they are consistent (the estimated probabilities sum to less than one, the log odds work out for all the groups, the results will not be affected by which group is the reference group, and we can do hypothesis testing across different groups).

This model depends on some assumptions:

- The model is correct: Observations are independent and multinomial, and the probabilities depend only on the predictors given.
- Independence of Irrelevant Alternatives: If a category is added or removed, it will not affect the relative odds of the other categories. (This can fail in *discrete choice models*, with the Red Bus/Blue Bus problem.) One can test for this using a Hausman test. This test removes one outcome and then compares the estimates for all other pairs to the original estimates; they should be approximately unchanged if IIA holds.

To visualize the data, boxplots of the independent variables for each outcome may again be useful.

Diagnostics are not well-defined (because each observation is associated with $K-1$ independent probabilities), but we may observe the estimated probabilities and use these for classification. Goodness of fit tests exist only when each $n_i > 1$. We may measure the usefulness of the model in classification as before, with C_{max} based on classifying everything into the single largest group (C_{pro} is the same).

If all the predictors are categorical, then the estimates for a logistic regression are identical to the estimates for a log-linear contingency table model with all possible interactions for the predictors.

4.4.2 Ordinal Multinomial Logistic Regression

In the case where outcomes are ordinal, it may be reasonable to treat the outcomes like numbers, particularly for large data sets or more complicated forms of analysis. However, predictions will not be integral, we cannot get the probability for each outcome (making classification harder), and such a model assumes that the outcomes are "equally spaced."

A *latent variable regression* or *ordinal logistic regression* might make more sense. Suppose Y^* is a continuous variable, and we observe $Y = j$ if $\alpha_{j-1} < Y^* \leq \alpha_j$, where $-\infty = \alpha_0 < \alpha_1 < ... < \alpha_J = \infty$. Ordinal logistic regression estimates the relationship between X and Y^* and the α_j. This regression assumes that $Y^*|X$ has a logistic distribution, while ordinal probit regression would assume a normal distribution. This is also called a *proportional odds model*, because it implies that:

$$L_j(x) = \ln\left(\frac{F_j(x)}{1-F_j(x)}\right) = \alpha_j + \beta_1 x_1 + ... + \beta_k x_k, j = 1, ...J-1$$

where $F_j(x) = P(Y \leq j|X)$, which are the *cumulative logits*. This model estimates $J-1$ constants and only one β. $L_j(x_l+1) - L_j(x_l) = \beta_l$ for all j, x_l, so that the odds of seeing a response below a given category are multiplied by e^{β_l} when x_l increases by one unit.

In this model, merging adjacent categories does not change the relationship; in simply removes one of the α_j. However, estimation with fewer categories is less efficient.

An alternative model is the *adjacent categories logit*. In this model, we assume that:

$$\log(\frac{p_{j+1}}{p_j}) = \alpha_j + \beta_1 x_1 + ... + \beta_k x_k$$

That is, if x_l increases by one unit, then the odds of being the the next category up are multiplied by e^{β_l}. Because β_l does not depend on the category, it also means that the log odds of being in the category m levels higher increase by $m\beta_l$.