Ty Darnell
Test 2 Notes

Chapter 3: Diagnostics

Potential Problems with the Regression Model

1. Regression function is not linear
2. Error terms do not have constant variance
3. Error terms are not independent
4. Model fits all but one or a few outlier observations
5. Error terms are not normally distributed
6. One or several important predictor variables have been omitted from model

Regression function not linear

- A nonlinear trend may be revealed by using a nonparametric regression technique such as LOWESS.
- The LOWESS algorithm estimates a curve that represents the main trend in the data, without assuming a specific mathematical relationship between Y and X.
- *If error variance appears constant only X needs to be transformed*
- If error variance does not appear constant, you may need to transform Y and or X.
- Transforming both X and Y to simultaneously linearize the relationship and normalize the distribution of errors is a **Box-Cox transformation**

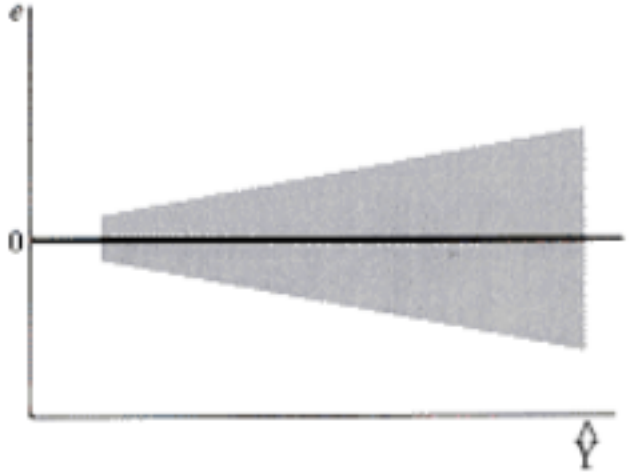
Linearity or Lack of Fit Test -test of whether the means of Y for the groups of replicates are significantly different from the fitted value on the regression line, using a kind of F test.

Error Variance not constant

- **Homoskedasticity**: σ^2 is constant over the entire range of X (as specified by OLS assumptions)
- **Heteroskedasticity**: σ^2 is not constant over the entire range of X (departure from OLS assumptions)

The regression model assumes homoskedasticity. Heteroskedasticity may look like a funnel or megaphone pattern (Figure 1). Can also be the reverse of Figure 1.

Figure 1: Heteroskedasticity



Residuals plotted against the predicted values

To detect Heteroskedasticity:

- Plot **absolute residuals or squared resid** against the predicted y and fit a lowess curve or a linear fit to help see a trend.
- **Brown-Forsythe Test** - Does not depend on normality of error terms. Requires user to break data into two groups and test for constancy error variance across groups.
- **Breusch-Pagan aka Cook-Weisberg Test** -Tests whether the log error variance increases or decreases linearly with the predictor(s). Requires large samples and assumes normal errors.

Variable Transformation to Equalize the Variance of Y (Remedy) - Box-Cox or Tukey's Ladder.

Weighted Least Squares (Remedy) - In weighted least squares observations are weighted in inverse proportion to the variance of the corresponding error term, so that observations with high variance are downweighted relative to observations with low variance.

Error Terms are not normally distributed

Look at Box Plot
Use Normal Probability (QQ) Plot of e
Use **Correlation Test for Normality**
Look for Kurtosis (fat tails) and skewness
Important to note: ***Standardizing a Variable Does Not Affect Shape of Distribution***
Tukey's ladder of powers spans a range of transformations to normalize the distribution of a variable in a data set.

Error Terms are not independent

errors can be autocorrelated or serially correlated

Figure 2: Error Terms are correlated

Durbin-Watson Test of Independence can be employed.

Chapter 5 Matrices

Multiply matrices
A * B results in a matrix with dimensions a rows and b columns.

Figure 3: Inverse of a matrix

$$1/(ad - bc)\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Figure 4: Matrix Representation in simple linear regression

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_3 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Figure 5: Multiple Linear Regression

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

Y and ϵ are the same as SLR

The Hat Matrix

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

Chapter 6 Multiple Regression

The interpretation of the parameters

1. β_0 indicates the mean of the distribution of Y when all of the independent variables equal 0.
2. β_k indicates the change in the mean response $E\{Y\}$ when X_k increases by one unit while all the other independent variables remain constant
3. σ^2 is the common variance of the distribution of Y

Sum of Squares

Figure 6: The sums of squares are defined identically in simple and multiple regression

$$SSTO = \Sigma(Y_i - \bar{Y})^2$$
$$SSE = \Sigma(Y_i - \hat{Y}_i)^2$$
$$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$$

with the relation SSTO = SSR + SSE

SSTO has n - 1 df
SSE has n - p df
SSR has p - 1 df

Mean squares are sums of squares divided by their respective degrees of freedom (df).
In particular, MSE = SSE/(n - p) is again the estimate of σ^2 , the common variance of ϵ and of Y.

F-Test

$MSR = \frac{SSR}{p-1}$

$MSE = \frac{SSE}{n-p}$

$F^* = \frac{MSR}{MSE}$

R^2 in SLR and MLR

Coefficient of Determination R^2 (SLR)
$R^2 = (SSTO - SSE)/SSTO = SSR/SSTO = 1 - SSE/SSTO$
The R^2 is interpreted as the proportion of the variation in y "explained" by the regression model.

Coefficient of Multiple Determination R^2 (MLR)
Defined the same as in SLR

Adjusted R-Square R_a^2
R_a^2 adjusts for the number of independent variables in the model (to correct the tendency of R^2 to always increase when independent variables are added to the model).
It is the percent of the variation in y that is explained by the set of independent variables included in the multiple regression.
$R_a^2 = 1 - ((n-1)/(n-p))(SSE/SSTO) = 1 - MSE/(SSTO/(n-1))$
It can be interpreted as 1 minus the ratio of the variance of the errors (MSE) to the variance of y, SSTO/(n-1).

t-test

Test Statistic $t^* = b_k/s\{b_k\}$

Standardized Regression Coefficients

$b_k^* = b_k(s(X_k)/s(Y))$
The standardized coefficient b_k^* measures the change in standard deviations of Y associated with an increase of one standard deviation of X.
Standardized coefficients permit comparisons of the relative strengths of the effects of different independent variables, measured in different metrics (= units).

Tolerance and Variance Inflation Factor

Tolerance (TOL)
R_k^2 is the R-square of the regression of X_k on the other p - 2 predictors in the regression and a constant.
Tolerance is between 0 and 1
TOL close to 1 means that R_k^2 is close to 0, indicating that X_k is not highly correlated with the other predictors in the model
TOL close to 0 means that X_k is highly correlated with the other predictors; one then says that X_k is collinear with the other predictors.
A common rule of thumb is that $TOL < .1$ This is an indication that collinearity may unduly influence the results.

VIF $(TOL)^{-1}$
VIF is the inverse of TOL
Large values of VIF therefore indicate a high level of collinearity.
The corresponding rule of thumb is that $VIF > 10$. This is an indication that collinearity may unduly influence the results.

GLM

The term general linear model is used for multiple regression models that include variables other than first powers of different predictors.

The X variables can also represent:

- different powers of a single variable (polynomial regression)
- interaction terms represented as the product of two or more variables
- qualitative (categorical) variables represented by one or more indicators (variables with values 1 or 0, aka "dummy variables")
- mathematical transformations of variables