Heteroscedasticity

Excessive **nonconstant variance**, or **heteroscedasticity**, can create technical difficulties with a multiple linear regression model. For example, if the residual variance increases with the fitted values, then prediction intervals will tend to be wider than they should be at low fitted values and narrower than they should be at high fitted values.

- We talked about perform transformation on y, but this doesn't often work very well.
- A generalization of **weighted least squares** is to allow the regression errors to be correlated with one another in addition to having different variances.
- In more extreme cases, we can use **generalized linear models** approach. Logistic regression is one of the GLM examples.

Test for heteroscedasticity: Breusch-Pagan Test

The basic idea is the variance of error should not change given different predictor values. If the equal variance assumption is not true, then the variance might be related to predictors.

- Fit the model $y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon$ and obtain the residuals e_is
- Build an auxiliary model (E) $e_i^2 = \gamma_0 + \gamma_1 X_{i,1} + \gamma_2 X_{i,2} + ... + \gamma_{p-1} X_{i,p-1} + \xi_i$
- Test the null hypothesis $H_0 : \gamma_1 = \gamma_2 = ... = \gamma_{p-1} = 0$ using the test statistic $LM = n * R^2$ from model (E).
- When H_0 is true, $LM \sim \chi^2(p)$
- In practice, it's also very common to use a F test for the Breusch-Pagan Test.

When heteroscedasticity exists: Weighted least squares

When we have $p - value < \alpha$ from the Breusch-Pagan Test, that suggests a nonequal variance from our dataset. Since OLSE doesn't hold under the violation of this assumption, we can use Weighted least squares instead to fit the model.

- Note: when we have OLSE $\hat{\beta}$, the estimate values stay the same, but the $Var(\hat{\beta})$ is inconsistent, it doesn't converge to the true variance of β and it's biased. If we perform the regular OLSE and ignore the heteroscedasticity, this would cause $se(\hat{\beta}_j)$ unreliable, and all the individual t tests have more false rejections.

Weighted least squares

- Now let's again consider $\mathbf{Y} = \mathbf{X}\beta + \epsilon$
- With the variance-covariance matrix of ϵ is

$$Var(\varepsilon) = \begin{pmatrix} \boxed{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \underline{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \underline{\sigma_n^2} \end{pmatrix} = W^{-1}$$

- If we define $w_i = 1/\sigma_i^2$, then let matrix W be a diagonal matrix containing these weights:

$$Var(A\varepsilon) = A\varepsilon A^T$$

$$W = \begin{pmatrix} \boxed{w_1} & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & w_n \end{pmatrix}$$

Weighted least squares

$$Y = X\beta + \varepsilon$$

$$\underbrace{W^{1/2} Y}_{\tilde{Y}} = \underbrace{W^{1/2} X}_{\tilde{X}} \beta + \underbrace{W^{1/2} \varepsilon}_{\tilde{\varepsilon}}$$

$$\boxed{\tilde{Y} = \tilde{X} \beta + \tilde{\varepsilon}}$$

$$\begin{aligned} \mathrm{Var}(\tilde{\varepsilon}) &= \mathrm{Var}(W^{1/2}\varepsilon) \\ &= W^{1/2}\,\mathrm{Var}(\varepsilon)\,W^{1/2} \\ &= W^{1/2}(W^{-1})W^{1/2} = I \end{aligned}$$

✓

- Now the least sum of squares is $SS = \sum w_i e_i^2$,
- $\hat{\beta}_{WLS} = (X^T W X)^{-1} X^T W y$ ← WLS
- With this setting, we can make a few observations:
  - Since each weight is inversely proportional to the error variance, it reflects the information in that observation. So, an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance (small weight).
  - The weights have to be known (or more usually estimated) up to a proportionality constant.

Multicollinearity

Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated with one another, or in a broader description, one of the columns of $X^T X$ is a linear combination of the others, i.e., $X^T X$ is close to non-invertible. **This problem always exists to some degree...**

- Consequences of multicollinearity:
  - The estimated regression coefficient of any one variable depends on which other predictors are included in the model.
  - The precision of the estimated regression coefficients decreases as more predictors are added to the model.
  - The marginal contribution of any one predictor variable in reducing the error sum of squares depends on which other predictors are already in the model.
  - Hypothesis tests for $\beta_i = 0$ may yield different conclusions depending on which predictors are in the model.

Symptoms of Multicollinearity and some solutions

Symptoms

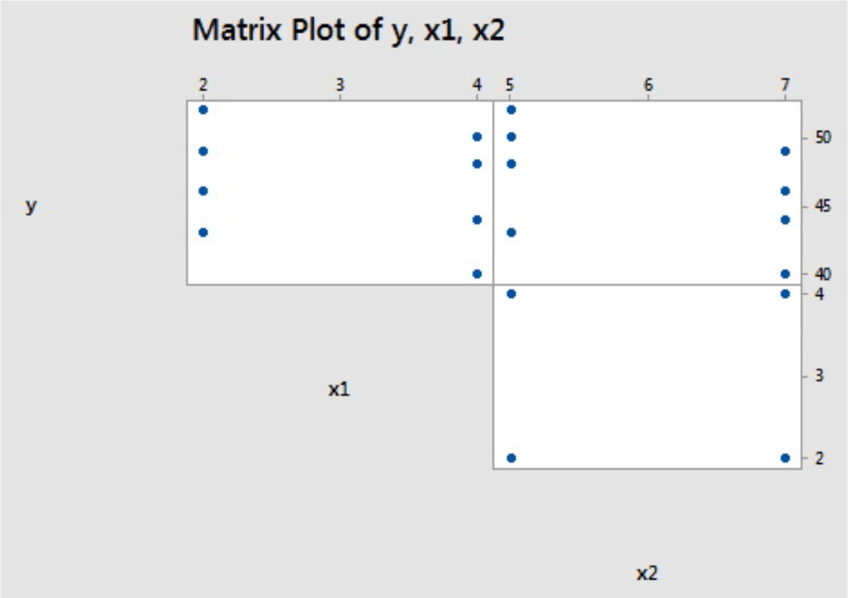
- There are only few coefficients show rejection/significance from t test, but F test is highly significant.
- When you add and drop variables, there are huge changes to the magnitude and sometimes even the sign of the remaining fitted coefficients.

Some general solutions

- If there is an unnecessary dummy, drop it.
- When adding X^2 or X^3 to the model and it's causing some problems, think about centering it first.
- When sample n is larger, it reduces some impact.
- Drop one or more obnoxious variables.
- Consider ridge regression.

Uncorrelated predictors

- In order to get a handle on this multicollinearity thing, let's first investigate the effects that uncorrelated predictors have on regression analyses.
- In the dataset below, we have $corr(X_1, X_2) = 0$.



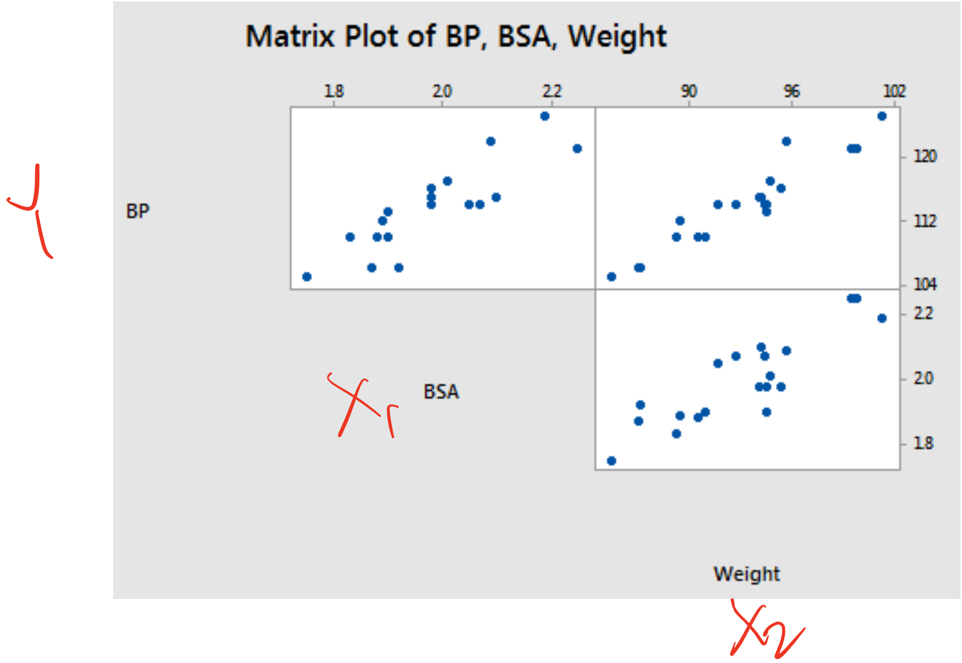
Uncorrelated predictors

- We can show a table summarizing when we include different predictors in the model how the output looks:

Model	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	t_s	SeqSSR
X_1 only	-1	1.47			-0.68	8.000
X_2 only			-1.75	1.35	-1.30	24.50
X_1, X_2	-1	1.41	-1.75	1.41	-0.71, -1.24	24.50
X_2, X_1	-1	1.41	-1.75	1.41	-1.24,-0.71	8.000

Highly correlated predictors

- Now let's take a look at a data set with highly correlated predictors. In the dataset below, y = BP and the predictors x1 = Weight and x2 = BSA, we have $corr(Weight, BSA) = 0.875$.



Highly correlated predictors

- We can show a table summarizing when we include different predictors in the model how the output looks:

Model	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$\hat{\beta}_2$	$se(\hat{\beta}_2)$	t_s	SeqSSR
X_1 only	1.201	0.093			12.92	505.5
X_2 only			34.44	4.69	7.34	419.9
X_1, X_2	1.039	0.193	5.83	6.06	5.39, 0.96	2.81
X_2, X_1	1.039	0.193	5.83	6.06	0.96,5.39	88.43

Detect Multicollinearity

$X_j \sim X_1 + X_2 + \cdots + X_{j+1} + \cdots + X_p$

R_j^2

- **Variance Inflation Factor (VIF)**: quantifies how much the variance is inflated. Recall that we learned previously that the standard errors — and hence the variances — of the estimated coefficients are "inflated" when multicollinearity exists. A variance inflation factor exists for each of the predictors in a multiple regression model. For example, the VIF_j for $\hat{\beta}_j$ is defined as $VIF_j = \boxed{\frac{1}{1-R_j^2}}$, where R_j^2 is the R^2-value obtained by regressing the jth predictor on the remaining predictors.
- A VIF of 1 means that there is no correlation among the jth predictor and the remaining predictor variables, and hence the variance of bj is not inflated at all. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of serious multicollinearity requiring correction.

VIF > 4 ⇒ warning ; VIF > 10 ⇒ cry

Interaction between quantitative predictors

$Y \sim X_1X_2$

Interaction terms between quantitative predictors allow the relationship between the response and one predictor to vary with the values of another predictor. Interestingly, this provides another way to introduce curvature into a multiple linear regression model.

- Typically, regression models that include interactions between quantitative predictors adhere to the **hierarchy principle**, which says that if your model includes an interaction term, X_1X_2, and is X_1X_2 is shown to be a statistically significant predictor of Y, then your model should also include the ”main effects,” X_1 and X_2, whether or not the coefficients for these main effects are significant.

If $\underline{X_1X_2}$ in model, include $\underline{X_1}$ & $\underline{X_2}$ also.

Overfitting

- When building a regression model, we don't want to include unimportant or irrelevant predictors whose presence can overcomplicate the model and increase our uncertainty about the magnitudes of the effects for the important predictors (particularly if some of those predictors are highly collinear).
- Such "overfitting" can occur the more complicated a model becomes and the more predictor variables, transformations, and interactions are added to a model. It is always prudent to apply a sanity check to any model being used to make decisions. Models should always make sense, preferably grounded in some kind of background theory or sensible expectation about the types of associations allowed between variables.
- Predictions from the model should also be reaosnable (over-complicated models can give quirky results that may not reflect reality).

Missing Data

Real-world datasets frequently contain missing values, so that we do not know the values of particular variables for some of the sample observations. For example, such values may be missing because they were impossible to obtain during data collection. Dealing with missing data is a challenging task. Missing data has the potential to adversely affect a regression analysis by reducing the total usable sample size. The best solution to this problem is to try extremely hard to avoid having missing data in the first place. When there are missing values that are impossible or too costly to avoid, one approach is to replace the missing values with plausible estimates, known as imputation. Another (easier) approach is to consider only models that contain predictors with no (or few) missing values. This may be unsatisfactory, however, because even a predictor variable with a large number of missing values can contain useful information.