Multicollinearity

- Read Section 7.5 in textbook.

- Multicollinearity occurs when two or more predictors in the model are *correlated* and provide *redundant information* about the response.

- Example of multicollinear predictors are height and weight of a person, years of education and income, and assessed value and square footage of a home.

- Consequences of high multicollinearity:

  1. Increased standard error of estimates of the β's (decreased reliability).
  2. Often confusing and misleading results.

Multicollinearity (cont'd)

- What do we mean by confusing results?
- Suppose that we fit a model with x_1 income and x_2 years of education as predictors and y intake of fruits and vegetables as the response.
- Years of education and income are correlated, and we expect a positive association of both with intake of fruits and vegetables.
- $t-$tests for the individual β_1, β_2 might suggest that none of the two predictors are significantly associated to y, while the $F-$test indicates that the model is useful for predicting y.
- Contradiction? No! This has to do with interpretation of regression coefficients.

Multicollinearity (cont'd)

- β_1 is the expected change in y due to x_1 given x_2 is already in the model.

- β_2 is the expected change in y due to x_2 given x_1 is already in the model.

- Since both x_1 and x_2 contribute redundant information about y once one of the predictors is in the model, the other one does not have much more to contribute.

- This is why the $F-$test indicates that at least one of the predictors is important yet the individual tests indicate that the contribution of the predictor, given that the other one has already been included, is not really important.

Detecting multicollinearity

- Easy way: compute correlations between all pairs of predictors. If some r are close to 1 or -1, remove one of the two correlated predictors from the model.

- Another way: calculate the **variance inflation factors** for each predictor x_j:
$$VIF_j = \frac{1}{1 - R_j^2},$$
where R_j^2 is the coefficient of determination of the model that includes all predictors $except$ the jth predictor.

- If $VIF_j \geq 10$ then there is a problem with multicollinearity.

- JMP: Right-click on *Parameter Estimates* table, then choose *Columns* and then choose *VIF*.

Multicollinearity - Example

- See Example 7.3, page 349. Response is carbon monoxide content of cigarettes and predictors are tar content (x_1), nicotine content (x_2) and weight (x_3).

- Estimated parameters in first order model: $\hat{y} = 3.2 + 0.96x_1 - 2.63x_2 - 0.13x_3$.

- $F = 78.98$ with $p-$value below 0.0001. Individual $t-$statistics and $p-$values: 3.37 (0.0007), -0.67 (0.51) and -0.03 (0.97).

- Note that signs on β_2 and β_3 are opposite of what is expected. Also, very high F would suggest more than just one signficant predictor.

- VIF were: 21.63, 21.89 and 1.33, so there is a multicollinearity problem. Correlations are $r_{12} = 0.98$, $r_{13} = 0.49$ and $r_{23} = 0.50$.

Multicollinearity - Solutions

- If interest is only in estimation and prediction, multicollinearity can be ignored since it does not affect $\hat{y}$ or its standard error (either $\hat{\sigma}_{\hat{y}}$ or $\hat{\sigma}_{y-\hat{y}}$).

- Above is true only if the x_p at which we want estimation or prediction is within the range of the data.

- If the wish is to establish association patters between y and the predictors, then analyst can:
  - Eliminate some predictors from the model.
  - Design an experiment in which the pattern of correlation is broken.

Multicollinearity - Polynomial model

- Multicollinearity is a problem in polynomial regression (with terms of second and higher order): x and x^2 tend to be highly correlated.

- A special solution in polynomial models is to use $z_i = x_i - \bar{x}_i$ instead of just x_i. That is, first subtract each predictor from its mean and then use the deviations in the model.

- Example: suppose that the model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ and we have a sample of size n.

- Let $\bar{x} = n^{-1} \sum_i x_i$ and define $z = x - \bar{x}$. Then fit the model

$$y = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + e.$$

Multicollinearity - Polynomial model (cont'd)

- Example: $x = 2, 3, 4, 5, 6$ and $x^2 = 4, 9, 16, 25, 36$. As x increases, so does x^2. $r_{x,x^2} = 0.98$.

- $\bar{x} = 4$ so then $z = -2, -1, 0, 1, 2$ and $z^2 = 4, 1, 0, 1, 4$. Thus, z and z^2 are no longer correlated. $r_{z,z^2} = 0$.

- We can get the estimates of the β's from the estimates of the γ's. Since

$$\begin{aligned} E(y) &= \gamma_0 + \gamma_1 z + \gamma_2 z^2 \\ &= \gamma_0 + \gamma_1 (x - \bar{x}) + \gamma_2 (x - \bar{x})^2 \\ &= (\gamma_0 + \gamma_1 \bar{x} + \gamma_2 (\bar{x})^2) + (\gamma_1 - \gamma_2 \bar{x}) x + \gamma_2 x^2. \end{aligned}$$

  Then: $\beta_0 = \gamma_0 + \gamma_1 \bar{x} + \gamma_2 (\bar{x})^2$, $\beta_1 = \gamma_1 - \gamma_2 \bar{x}$ and $\beta_2 = \gamma_2$.

Analysis of residuals - Chapter 8

- Residual analysis is important to detect departures from model assumptions ($\epsilon \sim N(0, \sigma^2)$ and observations are independent) and to determine lack of fit of the model.

- Estimated residuals are obtained as

$$e_i = y_i - \hat{y}_i.$$

- A *standardized residual* is defined as $e_i^* = e_i / RMSE$.

- Very roughly, we expect that standardized residuals will be distributed as standard normal random variables $N(0, 1)$ as long as the assumptions of the model hold.

Residual plots for lack of fit

- Scatter plots of residuals e_i against each of the predictors and against $\hat{y}$ provide information about lack of fit.

- Look for trends and changes in variability that indicate that there is some systematic (rather than purely random) effect on residuals.

- Look for no more than 5% residuals falling outside of $0 \pm RMSE$.

- Figure 8.4 is residuals against predictor when a second order model is fit to the data. No patterns are observable.

- Figure 8.3: when a first order model is fitted, there is a curvilinear pattern to residuals, indicating that predictor needs to be added to model in quadratic form.

Residual plots for unequal variances

- When the variance of ϵ_i depends on the value of $\hat{y}_i$, we say that the variances are $heteroscedastic$ (unequal).

- Heteroscedasticity may occur for many reasons, but typically occurs when responses are not normally distributed or when the errors are not additive in the model.

- Examples of non-normal responses are:
  - Poisson counts: number of sick days taken per person/month, or number of traffic crashes per intersection/year.
  - Binomial proportions: proportion of felons who are repeat offenders or proportion of consumers who prefer low carb foods.

- See Figures 8.9 and 8.10 for an example of each.

Residual plots for unequal variances

- Unequal variances also occur when errors are multiplicative rather than additive:
$$y = E(y)\epsilon.$$

- In multiplicative models, the variance of y is proportional to the squared mean of y:
$$Var(y) = [E(y)]^2\sigma^2,$$
so the variance of y (or equivalently, the variance of ϵ) increases when the mean response increases.

- See Figure 8.11 for an example of this.

Solutions for unequal variances

- Unequal variances can often be ameliorated by transforming the data:

Type of response	Appropriate transformation
Poisson counts	$\sqrt{y}$
Binomial proportions	$\sin^{-1}\sqrt{y}$
Multiplicative errors	$\log(y)$

- Fit the model to the transformed response.

- Typically, transformed data will satisfy the assumption of homoscedasticity.

Example

- Traffic engineers wish to assess association between $\#$ of crashes and daily entering vehicles at intersections.

- Data are $\#$ of crashes per year at 50 intersections and average (over the year) DEV.

- See JMP example on handout with results from fitting the simple model: crashes $= \beta_0 + \beta_1$ DEV.

- Plot of $\hat{\epsilon}$ against $\hat{y}$: as $\hat{y}$ increases so does spread of residuals around zero: indication that σ^2 is not constant for all observations.

- The same model was fitted, but this time using $\sqrt{\text{crashes}}$ as response. See new residual plot with no evidence of heteroscedasticity.

Checking normality assumption

- Even though we assume that $\epsilon \sim N(0, \sigma^2)$, moderate departures from the assumption of normality do not have a big effect on the reliability of hypothesis tests or accuracy of confidence intervals.

- There are several statistical tests to check normality, but they are reliable only in very large samples.

- Easiest type of check is graphical:
  - Fit model and estimate residuals $e_i = y_i - \hat{y}_i$.
  - Get a histogram of residuals and see if they look normal.

Checking normality - Q-Q plots

- We can also obtain a Q-Q (Quantile-Quantile) plot, also called Normal Probability Plot of the residuals.

- A QQ plot is the following:
  - On the y-axis: the expected value of each residual under the assumption that the residual in fact comes from a normal distribution. These are called $normal\ scores$.
  - On the x-axis: the estimated residuals.

- If residuals in fact are normal, then the plot will show a straight, 45^o line. Departures from the straight line indicate departures from normality.

Checking normality - Q-Q plots

- JMP will construct normal probability plots directly: Choose *Distribution*, then right-click on variable name and choose *Normal Quantile Plot*.

- A formula to compute normal scores and then construct plot by hand is given on page 394 - box. Not required learning.

- See Figures 8.20 (no departures from normality) and Exercise 8.18 on page 396 (again no departures).

- See figures on next page: simulated Normal and non-normal data (non-normal simulated by taking fourth power of normal data).

Outliers and influential observations

- Observations with large values of the residual $e_i = y_i - \hat{y}_i$ are called *outliers*.

- What is a large residual?

  - A residual e_i is large if it is outside of $0 \pm 3 \times RMSE$.
  - A standardized residual is large if it is outside of 0 ± 3.

- A *studentized residual* is similar to a standardized residual, but it is computed as
$$z_i^* = \frac{e_i}{RMSE\sqrt{1 - h_i}},$$
where h_i is the *leverage* of the ith observation. The higher h_i, the higher the studentized residual. JMP gives the z_i^*, we do not need to compute h_i.

Outliers and influential obs - Cook's D

- Cook's D measures the *influence* of each observation on the value of the estimated parameters.

- Definition:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)MSE} \left[\frac{h_i}{(1-h_i)^2} \right].$$

- Cook's D is large for the ith observation if the residual $(y_i - \hat{y}_i)$ is large or the leverage h_i is large or both.

- Although not obvious from the formula, D_i measures the difference in the estimated parameters in a model fitted to all the observations and an model fitted to the sample that omits y_i. If D_i is small, then the ith observation does not have an overwhelming effect on parameter estimates.

Cook's D

- We can compute D_i for each of the n sample observations using JMP (or SAS).

- After fitting the model:
  - Right-click on *Response* icon
  - Choose *Save Columns*
  - Choose *Cook's D Influence*

- How big is big? Compare D_i to the other D_j in the sample. Notice whether any of the D_i are substantially bigger than the rest. If so, those observations are influential.

Outliers and influential obs - Example

- Where do outliers come from? Measurement or recording error or if no error, perhaps inappropriate model.

- Example FastFood, Table 8.5 in book. Response y is sales in 1000$ for fastfood outlets in four cities. Predictors are traffic flow (x_4, in 1000 cars) and three dummies (x_1, x_2, x_3) to represent the four cities.

- Suppose that observation 13 in city 3 for which traffic flow is 75,800 cars had sales recorded as $82,000 rather than $8,200 (someone put the decimal in the wrong place) and fit a first order model

$$E(y) = \beta_0 + \beta_1 x_x + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

Outliers and influential obs - Example

- Estimated prediction equation is: $\hat{y} = -16.46 + 1.1x_1 + 6.14x_2 + 14.5x_3 + 0.36x_4$.
- $RMSE = 14.86$ and $CV = 163.8$.
- $F-$test for utility of model is only 1.66 with a $p-$value of 0.2, meaning that none of the predictors are useful for predicting sales.
- $R_a^2 = 0.10$: only about 10% of the observed variability in sales can be attributable to traffic flow and city.
- Results are not what was expected. Traffic flow is expected to be associated to fast food sales.

Outliers and influential obs - Example

- Residual plots: observation 13 is clearly an outlier (see next page).

- Standardized residual is 4.36, outside of expected interval (-3, 3) and ordinary residual is 56.46, outside of $0 \pm 3 \times 14.86$.

- Cook's D for obs 13 is 1.196, about 10 times larger than the next largest D in the sample.

- If now we correct the mistake and use $8,200, results change dramatically: $F-$statistic is 222.17 with a very small $p-$value, $R_a^2 = 97.5\%$ and largest standardized residual is only -2.38.

- One outlier can have an enormous effect on the results! Important to do a residual analysis.