Multiple Linear Regression

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

What is Linear Regression?

- It is a hypothetical model of the relationship between two variables.
    - The model used is a linear one.
- Theoretical assumption:
    - For every one unit of change in the independent variable there will be a consistent and uniform change in the dependent variable.
    - Therefore, we describe the relationship using the equation of a straight line.
    - This can be seen as a way of predicting the value of one variable from another.

Before Regression

- We need to establish evidence to support going ahead with building a predictive model.
- If we are asserting a relationship,
  - We need to investigate if there is any evidence of a relationship using the appropriate test and make a decision based on the results (strength, direction etc.).
- If we are asserting a differential effect for different groups,
  - We need to investigate if there is any difference using the appropriate test and make a decision based on the result.

Regression - what does it allow you do?

- Questions it allows you to answer:
  - How well a set of variables is able to predict an outcome variable?
  - Which variable in a set is the best predictor?
  - Whether a variable is still able to predict an outcome when controlling for a particular variables.
- Prediction:
  - Really what we are looking at is the variance in the outcome variable and how much of the variance could be considered to be explained by the predictor variables.

Parametric v Non-Parametric Tests

Testing for	**Parametric**	**Non-Parametric**
Correlation	Pearson	Spearman/Kendall
Comparing Groups - 2 independent samples scale data	Independent t-test	Mann Witney U
Comparing Groups - 2 related (paired) samples scale data	Paired Samples t-test	Wilcoxon Signed Rank
Comparing Groups - 2 independent samples categorical data		Chi-Square
Comparing Groups - repeated measures categorical data		McNemar's
Comparing Groups - more than 2 independent samples	One Way ANOVA	Kruskal Wallis

Simple Linear Regression

Four important statistics:

- F statistic:
  - Whether the model as a whole predicts the dependent variable.
  - Its statistical significance is the significance of the model.
- Regression coefficients (Beta values):
  - Measure the strength and direction of relationships between independent variables and the dependent variance.
- Significance scores for the regression coefficients:
  - Tell us whether the contribution of each variable is statistically significant.
- R^2 Statistic or Adjusted R^2 Statistic:
  - Measures the model's overall predictive power and the extent to which the variables explain the variation found in the dependent variable.

How to Interpret Beta Values?

- Beta values:
  - the change in the outcome associated with a unit change in the predictor.
- Standardised beta values:
  - tell us the same but expressed as standard deviations.
- If we have standardised our variables in advance this will be virtually the same.

R^2 vs Adjusted R^2

- Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases.
  - Therefore, a model with more terms may appear to have a better fit simply because it has more terms.
- If a model has too many predictors it begins to model the random noise in the data.
  - This condition is known as overfitting the model and it produces misleadingly high R-squared values and a lessened ability to make predictions.
- Adjusted R^2:
  - Adjusts value of R2 based on number of variables in the model.
  - Report Adjusted R2 for multiple linear regression.

What is Multiple Regression?

- Simple linear Regression is a model to predict the value of one variable from another.
- Multiple Regression is a natural extension of this model:
    - We use it to predict values of an outcome from several predictors.
    - It is a hypothetical model of the relationship between several variables.

Multiple Regression as an Equation

- With multiple regression the relationship is described using a variation of the equation of a straight line.

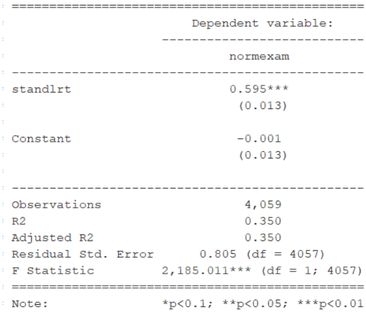
$$\boxed{y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_n X_n + \epsilon_i}$$

- b_0 is the intercept.
  - The intercept is the value of the Y variable when all Xs = 0.
  - This is the point at which the regression plane crosses the Y-axis (vertical).
- b_1 to b_n are the regression coefficient for variable 1 to n.

Conducting Linear Regression

Use lm to create our model and assign it to a variable:

```
model1 = lm(regression$normexam~regression$standlrt)
#Tidy output of all the required stats
stargazer(model1, type="text")
```



```
===============================================
                        Dependent variable:
                    ---------------------------
                             normexam
-----------------------------------------------
standlrt                     0.595***
                              (0.013)

Constant                      -0.001
                              (0.013)

-----------------------------------------------
Observations                   4,059
R2                             0.350
Adjusted R2                    0.350
Residual Std. Error      0.805 (df = 4057)
F Statistic         2,185.011*** (df = 1; 4057)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Dummy variables

- Many variables we are interested in using as predictors are categorical.
    - E.g. we are interested in the effect gender or religion or income category has.
- Because they have no scale it makes no sense to think of the effect of a unit of increase in these as it would for a continuous variable.
- But we can think in terms of the differential effect for groupings within the category.
- We can transform the categorical variable into a series of dummy variables which indicate whether a particular case has that particular characteristic.
- Dummy variables may also be referred to as indicator variables.

Our regression dataset (Regression.sav)

- Previously looked at standlrt as a predictor of normexam.
- We might also be interested in gender and whether it has an influence.
  - There is significant research that gender has an influence in educational achievement.
- We might also be interested in exploring the type of school a student attends (either mixed-sex or single-sex boys or girls schools) as this might also have an effect on a student's examination performance ('normexam').
- We are interested in exploring if there is a differential effect for students of different genders and for students of different genders attending different types of school.

Dummy Variables

- If we just added them into the model as they are what would happen?
- The values of these two variables would be treated as *real numerical values* rather than just arbitrary numbers representing specific categories.
- So we need to transform these into *dummy variables* before we can add them into the regression model.

Dummy variables

- Recode to 0 (reference category) and 1 (category of interest).
- Aim is to explore if there is a differential effect for the category of interest when compared to the reference category.
- Indicator variable:
  - Switch effect ON (1) or OFF (0).
- Before including in the regression model need to first establish if this makes sense to include as a predictor.
  - Investigate using an independent t-test.

Recoding Variables

```
library(car)
regression$gender=recode(regression$GIRL, '0=1;1=2')
```

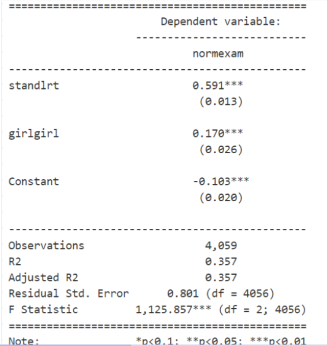
(this creates a variable gender which recodes GIRL, if 0 gender is set to 1 and if 1 gender is set to 2)

- This assumes that you have called your dataset regression.
- **This is already done in the dataset the dataset regression we are using**

Second Model including gender

```
model2<-lm(regression$normexam~
    regression$standlrt+regression$girl)
#Tidy output of all the required stats
stargazer(model2, type="text")
```

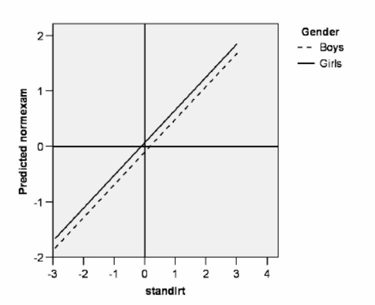
Variable girl – value girl



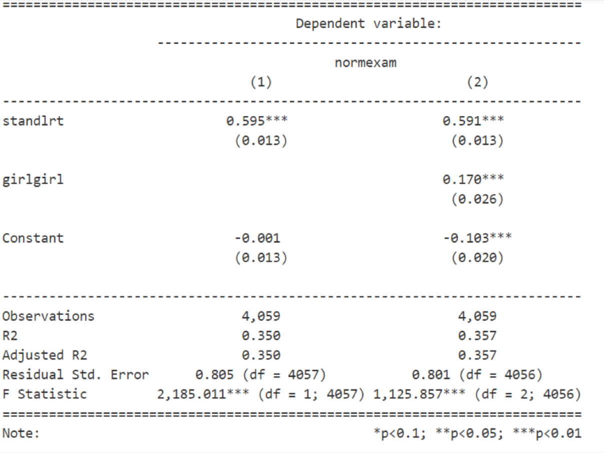
```
===============================================
                        Dependent variable:
                    ---------------------------
                             normexam
-----------------------------------------------
standlrt                     0.591***
                              (0.013)


girlgirl                     0.170***
                              (0.026)


Constant                    -0.103***
                              (0.020)


-----------------------------------------------
Observations                   4,059
R2                             0.357
Adjusted R2                    0.357
Residual Std. Error      0.801 (df = 4056)
F Statistic         1,125.857*** (df = 2; 4056)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Example

- In effect the consequence of adding the dummy variable 'girl' to the model is to create two lines of best fit that have the same gradient (0.591) but different intercepts (constants) (i.e. -0.103 for boys and 0.067 for girls).
- In other words, and as illustrated below, this model can be represented as two parallel lines with the vertical distance between both lines being 0.170:



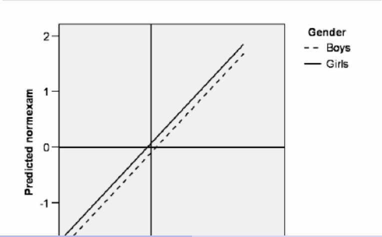
Comparison of models



```
===========================================================================
                                      Dependent variable:
                    -------------------------------------------------------
                                           normexam
                                (1)                         (2)
---------------------------------------------------------------------------
standlrt                     0.595***                    0.591***
                              (0.013)                     (0.013)

girlgirl                                                 0.170***
                                                          (0.026)

Constant                      -0.001                     -0.103***
                              (0.013)                     (0.020)


---------------------------------------------------------------------------
Observations                   4,059                       4,059
R2                             0.350                       0.357
Adjusted R2                    0.350                       0.357
Residual Std. Error         0.805 (df = 4057)           0.801 (df = 4056)
F Statistic         2,185.011*** (df = 1; 4057) 1,125.857*** (df = 2; 4056)
===========================================================================
Note:                                           *p<0.1; **p<0.05; ***p<0.01
```

Example

- This model is fine but it does assume that the gap in exam scores at age 16 between boys and girls remains constant (at 0.170 points) regardless of a student's prior educational achievement at 11 ('standlrt').
- But it may be that high achieving boys at age 11 could be outperforming girls at age 16 and/or that low performing boys at age 11 may fall even further behind their female counterparts by the age of 16?
  - We are hypothesising here, in effect, that the gradients of both these lines of best fit might not be the same and thus these two lines might not actually be parallel.
- So to really improve this we would need to look at including an interaction term

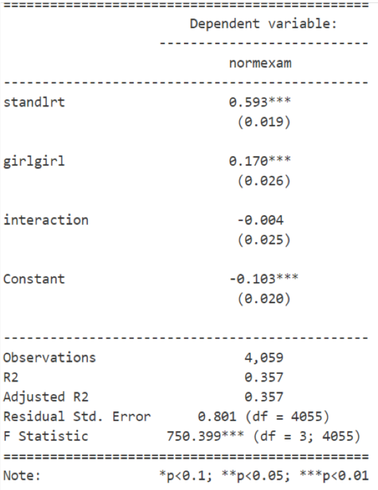


Interaction term

- We can test this hypothesis by adding what is referred to as an 'interaction term' to the model that is basically a new variable calculated by multiplying 'standlrt' by 'girl'.
- First, therefore, we need to create a new variable (which we can call 'standlrt_girl').

```
regression$interaction <- regression$standlrt *
as.numeric(regression$girl)

*Already in our dataset
```

Model 3 – including an interaction term



```
===============================================
                        Dependent variable:
                    ---------------------------
                             normexam
-----------------------------------------------
standlrt                     0.593***
                             (0.019)

girlgirl                     0.170***
                             (0.026)

interaction                   -0.004
                             (0.025)

Constant                     -0.103***
                             (0.020)

-----------------------------------------------
Observations                   4,059
R2                             0.357
Adjusted R2                    0.357
Residual Std. Error      0.801 (df = 4055)
F Statistic         750.399*** (df = 3; 4055)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

$Predicted\ normexam =$
$-.103 + 0.593 * standlrt + 0.17 *$
$girl - 0.0 * interaction$

$Boys\ normexam =$
$-.103 + 0.593 = 0.49$
$Girls\ normexam =$
$-.103 + (0.593 - 0.004)$
$standlrt + .17 =$
$-.103 + .589 * standrt + .17$

Conclusion that there is no real difference for low/high performing students of different gender.

Interaction term

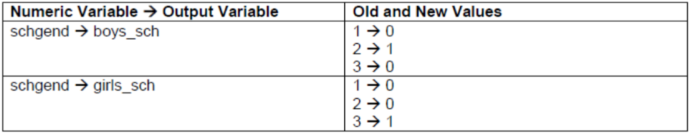
- We can see how the inclusion of the interaction term has basically had the consequence of giving us two lines of best fit (one for boys and one for girls) that now having differing intercepts as well as differing gradients or slopes.
- On this occasion we already knew that the difference in the slopes was minor (and not statistically significant) and this is confirmed by these two lines of best fit where the difference in the gradients of both lines (0.593 compared to 0.589) is marginal and wouldn't even be obvious to the eye if the two lines were plotted on a graph.
- We would conclude that this interaction term has no significant effect on the model.

Dummy variables with more than two categories

- We need to create a number of dummy coded variables (the number of which will be one less than the number of categories in the original variable).
- We need to define one of the categories as the reference category.
  - Which category you pick is entirely up to you but it should make sense in the context of your question.
  - For ordinal variables, it is usually best to select the lowest-ranked category as the reference category.

Dummy Variables

```
regression$boy_sch = ifelse(regression$schgend=="boysch", 1, 0)
regression$girl_sch = ifelse(regression$schgend=="girlsch", 1, 0)
```

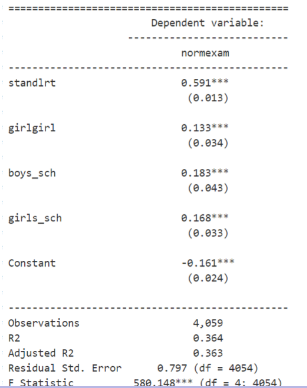


Numeric Variable → Output Variable	Old and New Values
schgend → boys_sch	1 → 0 2 → 1 3 → 0
schgend → girls_sch	1 → 0 2 → 0 3 → 1

- This has already been done in the regression dataset we are using.
- It is always worth just checking that you have recoded things properly by comparing the old and new variables via simple crosstabs.

Model 4 – including school gender

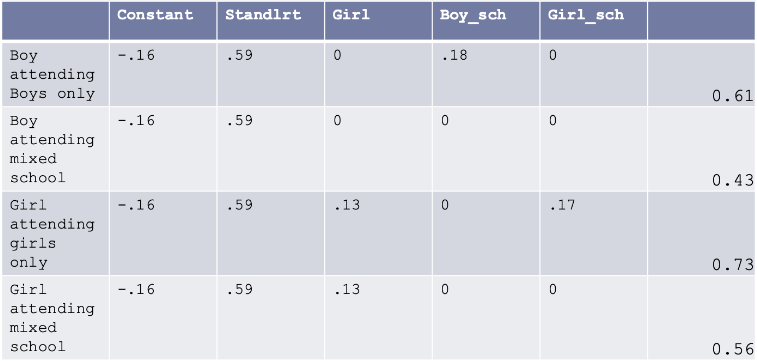
```
model3<−lm(regression$normexam~regression$standlrt+
regression$girl+regression$boys_sch+regression$girls_sch)
Stargazer(model3, type="text")
```



```
===============================================
                        Dependent variable:
                    ---------------------------
                             normexam
-----------------------------------------------
standlrt                     0.591***
                              (0.013)

girlgirl                     0.133***
                              (0.034)

boys_sch                     0.183***
                              (0.043)

girls_sch                    0.168***
                              (0.033)

Constant                    -0.161***
                              (0.024)

-----------------------------------------------
Observations                   4,059
R2                             0.364
Adjusted R2                    0.363
Residual Std. Error      0.797 (df = 4054)
F Statistic          580.148*** (df = 4; 4054)
```

Interpreting the model

```
Predicted normexam=−.16+.59∗standlrt+0.13∗girl+0.18∗boy_sch
    +0.17∗girl_sch
```



	Constant	Standlrt	Girl	Boy_sch	Girl_sch	
Boy attending Boys only	-.16	.59	0	.18	0	0.61
Boy attending mixed school	-.16	.59	0	0	0	0.43
Girl attending girls only	-.16	.59	.13	0	.17	0.73
Girl attending mixed school	-.16	.59	.13	0	0	0.56

Model 4

So we can see a difference in scores for boys attending boys only and girls attending girls only.

Reflecting on Regression

R and R^2

- R: The correlation between the observed values of the outcome, and the values predicted by the model.
- R^2: The proportion of variance accounted for by the model.
- $Adj.\ R^2$: An estimate of R2 in the population (shrinkage).

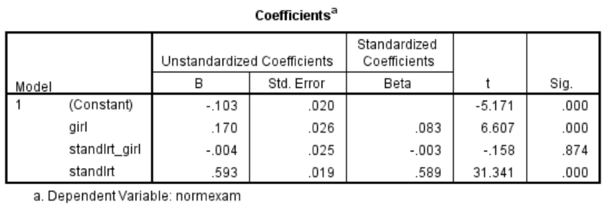
Analysis of Variance: ANOVA

The F-test:

- Looks at whether the variance explained by the model is better than one with no predictors.
- It tells us whether using the regression model is significantly better at predicting values of the outcome than using the mean.

Output: betas

- Beta values: the change in the outcome associated with a unit change in the predictor.
- Standardised beta values: tell us the same but expressed as standard deviations.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.103	.020		-5.171	.000
	girl	.170	.026	.083	6.607	.000
	standlrt_girl	-.004	.025	-.003	-.158	.874
	standlrt	.593	.019	.589	31.341	.000

a. Dependent Variable: normexam

Concerns for MLR - Choosing variables for MLR

- Have a sound theoretical basis for including your variables.
- Explore the data using univariate and bivariate analysis.
- Only include variables that have potentially informative results or which serve as controls.
- In large models:
  - Introduce variables sequentially.
  - Start by introducing small groups.
  - As you include more variables note not only how they operate but how the scores for other variables and their statistical significance change as well.
  - If the model is statistically stable, these will stay within the same general range.
  - If the values start changing a lot, it suggests you need to conduct some further exploratory work before trusting your findings.

Concerns for MLR - Degrees of Freedom

- The number of observations 'free to vary' in a sample.
- Every observation increases the df by one.
- But every coefficient the model estimates decreases it by one.
- Thus the more variables included in the model the more the df lowers which reduces the tests ability to find a statistically significant result.

Why would this be a problem?

- Reduces our statistical power.
- Increase possibility of Type II error.

Residuals

- Coefficients and statistical significance are calculated on the assumption that a straight line is a good model of the relationship.
- How well this line serves as a model of this relationship can be checked by looking at how well the observed data sit in relation to predicted values along this line.
    - Measured by the vertical distance from the prediction to the actual observation - the residual.
- Regression coefficients are calculated so that the resulting line has the lowest possible accumulation of residuals, minimising the overall distance between the observation and the predictions.

Regression Diagnostics

Methods for determining whether a regression model fit to data adequately represents the data:

- Unusual data in linear models,
- Non-normality,
- Non-constant error variance,
- Nonlinearity in linear models.

Some Terms

- Outliers: An observation that has a large residual i.e. the observed value for the point is very different from that predicted by the regression model.
- Leverage points: An observation that has a value of x that is far away from the mean of x.
- Influential observations:
    - An observation that changes the slope of the line.
    - They have a large influence on the fit of the model.

Concerns for MLR - Outliers (residuals)

- MLR can be greatly influenced by outliers.
  - They 'pull' the equation away from the general pattern.
- Sometimes it is reasonable to delete the outliers.
- Sometimes you need to retain the outlier.
  - it may be your most important observation.
- Sometimes it is a good idea to do your analysis twice – once with and once without the outlier to see how much influence it is having on the model.
- If you remove outliers, do repeat this residual analysis again, you may have introduced new outliers.

Why would this be a problem?

Bias and increased possibility of Type I error.

Concerns for MLR - Outliers

- Look at the Minimum and Maximum values of Std. Residual (Standardised Residual) subheading.
  - We would expect 95% to fall within +1.96 and -1.96 if our p value cut-off is 0.05 (5%).
  - If the minimum value is equal or below -3.29, or the maximum value is equal or above 3.29 (and out dataset has more than 80 cases) then you have outliers.
- Cook's distance:
  - Measures the effect of deleting a case.
  - Look at Cook's distance for values greater than one.
- Should also graphically inspect relevant plots.

Why would this be a problem?

Bias and increased possibility of Type I error.

Concerns for MLR - Constant variation

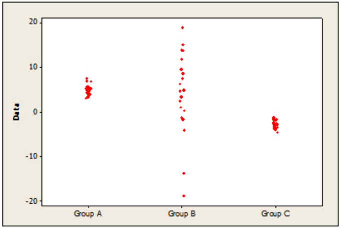
- Sometimes the regression line is not well suited to the data.
  - E.g. we may get a fan effect indicating that the regression line is better at predicting low values of the dependent variable than it is at predicting high values (or vice versa).
- It is not appropriate therefore to use MLR in these cases even if there is strong coefficients and statistical significance.
- In a well-fitting regression model the variation around the line is constant along the line.

Why would this be a problem?

Bias and increased possibility of Type I error.

Homoscedasticity

- Having the same scatter.
- Having data values that are scattered, or spread out, to about the same extent.



Group A and Group C exhibit homoscedasticity.

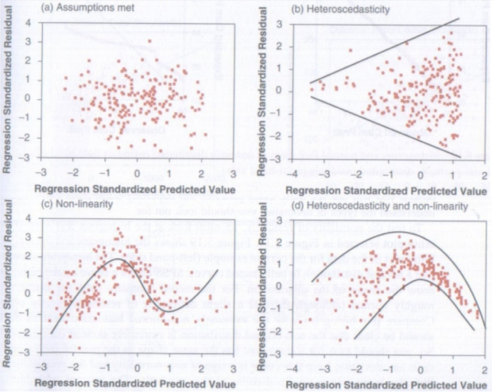
Group A and Group B exhibit heteroscedasticity (literally, "different scatter")

What about Groups B and C?

Why would this be a problem?

Bias and increased possibility of Type I error if heteroscedasticity.

Homoscedasticity: ZRESID vs. ZPRED



From Andy Field:

You are looking for a plot like the top left.

Concerns for MLR - Normality of residuals

- Should follow a normal distribution with a mean of 0.
- Since the residuals measure where the points fall in relation to the regression line, a normal distribution of residuals indicates that the same number of points fall above and below the line.
- A residual of 0 means a point is on the line, a mean of 0 means that the line is in the middle of the points.
- The bell shape means that most are close to the line and there are fewer father from it.

Why would this be a problem?

Bias and increased possibility of Type I error if heteroscedasticity.

Normality of Errors: Histograms and P-P plots

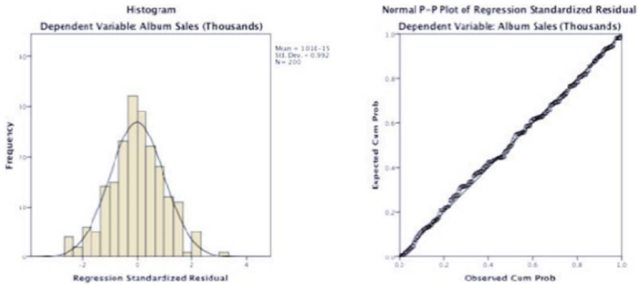


FIGURE 8.24 Histograms and normal P-P plots of normally distributed residuals (left-hand side) and non-normally distributed residuals (right-hand side)

Concerns for MLR - Collinearity

- Occurs when two or more independent variables contain strongly redundant information.
- If variables are collinear then it means there is not enough distinct information in these variables for MLR to operate - they are essentially measuring the same thing.
- If we conduct MLR with collinear variables then the model will produce invalid results.
- Need to check for collinearity by examining a correlation matrix that compares your independent variables with each other.
- A correlation coefficient above 0.8 suggests collinearity **might** be present.
    - If it is you need to analyse your variables in separate regression equations and then examine how they operate together.

Why would this be a problem?

Bias and increased possibility of Type I error if heteroscedasticity.

Collinearity

What can we do?

- Delete all but one of the collinear variables from the model.
- The one remaining is a proxy for the underlying concept.
- Combine them into an index mathematically (e.g. multiplying, adding etc.) - must make sense theoretically.
- Estimate a latent variable using principal component analysis or factor analysis.

Generating the relevant plots and statistics

In R

Details of residuals are included in the output of summary:

```
summary(model1)
#You can get the major plots
#using the plot function
plot(model)
```

Will give you:

- Residuals vs. fitted values
- Q-Q plots
- Scale Location plots
- Cook's distance plots.

Normality of Residuals in R

```
#Create histogram
resid(model) #List of residuals

#A density plot of the residuals
plot(density(resid(model)))

#Create a QQ plot
qqPlot(model, main="QQ␣Plot")
#qq plot for studentized resid
leveragePlots(model) # leverage plots
```

Collinearity

- VIF = variance inflation factors.
- Tells you what the standard error would be if the variable was uncorrelated with others.
- Values of VIF that exceed 2.5 should be cause for concern.
- Tolerance - should be concerned if it is lower than 0.4.

(Roger Tarling Statistical Modelling for Social Researchers).

```
library(car)
#Collinearity
vifmodel<-vif(model1)
#Calculate tolerance
1/vifmodel
```

Influential Outliers

- **Cook's distance** measures the effect of deleting a given observation.
- Points with a large **Cook's distance** are considered to merit closer examination in the analysis.
- Investigate any point over 4/n, where n is the number of observations.
- You can choose to eliminate those with influential values.

```
cooks<-cooks.distance(model)
plot(cooks.distance(model1), ylab="Cook's statistic")
```

You can plot and analyse this.

Concerns for MLR - Causality

- Need to take care about asserting causality on the basis of your analysis.
- Causality:
  - Association
  - Time order
  - Non-spuriousness
- Regression can reveal associations but they do not document time order.
- By including other factors we can control for spurious effects.
- But there is always the possibility that there is an untested spurious factor.

Generalization

- When we run regression, we hope to be able to generalize the sample model to the entire population.
- To do this, several assumptions must be met.
- Violating these assumptions stops us generalizing conclusions to our target population.

Straightforward Assumptions

- Variable Type:
  - Outcome must be continuous (Note: does not have to be normal but will make your life easier if it is).
  - Predictors can be continuous or dichotomous.
- Non-Zero Variance:
  - Predictors must not have zero variance.
- Linearity:
  - The relationship we model is, in reality, linear.
- Independence:
  - All values of the outcome should come from a different person.

The More Tricky Assumptions

- No Multicollinearity: Predictors must not be highly correlated.
- Homoscedasticity/Constant variance: For each value of the predictors the variance of the error term should be constant.
- Normally-distributed Errors.

Another Assumption

Independence of errors is also an assumption for time series data (beyond the scope of this module).

How to conduct an MLR

- Research and develop a theory.
- Identify your variables and scales needed.
- Collect your data.
- Screen your data and make decisions about.
    - Representativeness.
    - Missing data.
    - Univariate Outliers.
    - Normality.

How to conduct and report MLR

- Make sure your data meets the assumptions required.
    - Outliers
    - Random normal distribution of errors
    - Homoscedasticity
    - Linearity
    - Collinearity of data
    - Non-zero variances
- If your data doesn't meet the assumptions?
    - Investigate why and whether a multiple regression is really the best way to analyse it.

How to conduct and report MLR

- Outliers
  - Look at the Minimum and Maximum values next to Std. Residual (Standardised Residual) subheading.
  - If the minimum value is equal or below -3.29, or the maximum value is equal or above 3.29 then you have outliers.
  - Look at Cook's distance for values greater than one (Descriptives and Frequencies)
- Note: If you remove outliers, do repeat this residual analysis again, you may have introduced new outliers.

Reporting this:

An analysis of standard residuals was carried out on the data to identify any outliers, which indicated two cases for concern which were deleted.

How to conduct and report MLR

- If no outliers found, report the min and max.
- An analysis of standard residuals was carried out, which showed that the data contained no outliers (Std. Residual Min = -1.90, Std. Residual Max = 1.70).

How to conduct and report MLR

- Collinearity
    - $library(car)$
    - $vif(model)$
    - $sqrt(vif(model))\#Checkif > 2.5$
- Independent errors
    - $library(car)$
    - $durbinWatsonTest(model)$

How to conduct and report MLR

Collinearity:

- Coefficients table.
- We are interested in Tolerance and VIF.
- If the VIF value is greater than 2.5 or the Tolerance is less than 0.4, then you have concerns over multicollinearity.

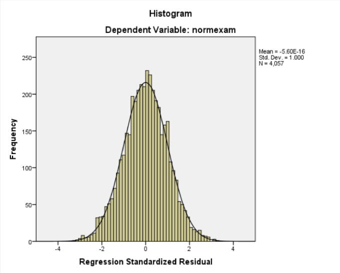
Otherwise, your data has met the assumption of collinearity and can be written up something like this:

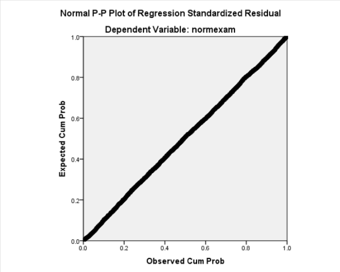
"Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Gender, Tolerance = .99, VIF = 1.00; Standard Reading Test, Tolerance = .45 VIF = 2.23, Interaction effect Standard Reading Test * Gender Tolerance = .45, VIF = 2.23)."

How to conduct and report MLR

- Normality of residuals, homoscedasticity.
- In R:
  - $plot(model)$

How to conduct and report MLR





Random Normal Distribution of Errors

Report as:
"The histogram of standardised residuals indicated that the data contained normally distributed errors, as did the normal P-P plot of standardised residuals, which showed all points were extremely close to the line"

How to conduct and report MLR

- **Non zero variances**.
- Check your descriptive statistics.

Reporting this:

"The data also met the assumption of non-zero variances (Exam Age 16 Variance = .995; Standard Reading Test Score Variance = .987 Gender= .240, Interaction Effect Standard Reading Score * gender = .542)"

How to report MLR

- Regression results are often best presented in a table.
- You should at least present the unstandardized or standardized slope (beta), whichever is more interpretable given the data, along with the t-test and the corresponding significance level.
- Tables are useful if you find that a paragraph has almost as many numbers as words.
- If you do use a table, do not also report the same information in the text. It's either one or the other.

How to report MLR - Example (CA Example)

A multiple regression analysis was conducted to determine if a student's score for well-being, mature student status and the type of educational institute being attended could predict a student's academic satisfaction.

In order to include the type of educational institute being attended in the regression model it was recorded into two variables iot_dummy (0 for university or college, 1 for IoT), and college_dummy (0 for university or IoT, 1 for college).

Examination of the histogram, normal P-P plot of standardised residuals and the scatterplot of the dependent variable, academic satisfaction, and standardised residuals has shown that some outliers existed. However, examination of the standardised residuals has shown that none could be considered to have undue influence (95% within limits of -1.96 to plus 1.96 and none with Cook's distance > 1 as outlined in Field (2013).

Examination for multicollinearity has shown that the tolerance and variance influence factor measures were within acceptable levels (tolerance > 0.4, VIF < 2.5) as outlined in Tarling (2008). The scatterplot of standardised residuals has shown that the data met the assumptions of homogeneity of variance and linearity. The data also meets the assumption of non-zero variances of the predictors.

Example: Julie Pallant's Survey.dat

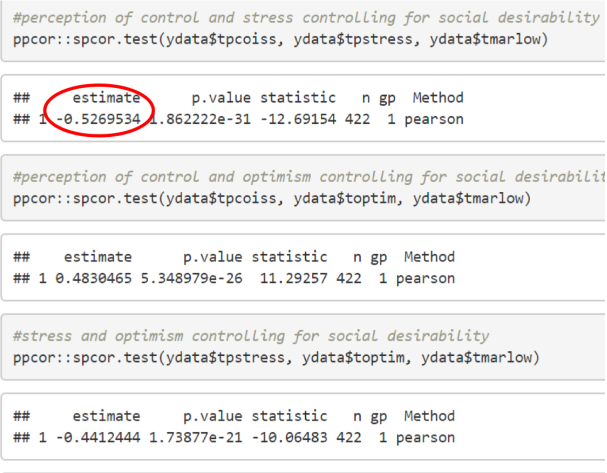
Partial Correlations

- Measures the relationship between two variables, **controlling** for the effect that a third variable has on them both.
- Usually a variable you suspect is influencing your two variables of interest.
- This can artificially inflate the correlation co-efficient found.

Partial Correlation

- Using survey.sav/survey.dat (Julie Pallant)
- Interested in exploring the relationship between scores on the Perceived Control of Internal States Scale (tpcoiss) and scores on the Perceived Stress Scale (tpstress) but controlling for what is known as *socially desirable responding bias* (tendency to present yourself in a socially desirable way).
- This tendency is measured using Marlowe-Crowne Social Desirability Scale (tmarlow).

Partial Correlation



```
#perception of control and stress controlling for social desirability
ppcor::spcor.test(ydata$tpcoiss, ydata$tpstress, ydata$tmarlow)
```

```
##     estimate      p.value statistic   n gp  Method
## 1 -0.5269534 1.862222e-31 -12.69154 422  1 pearson
```

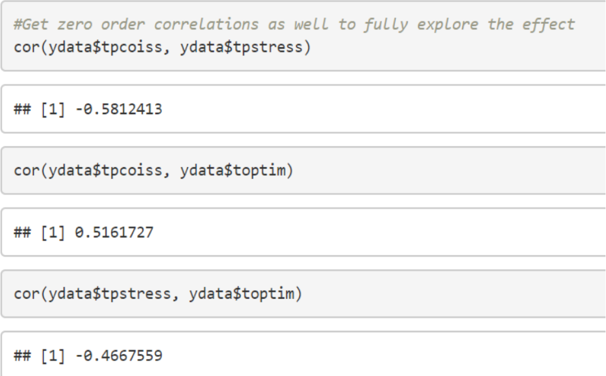
```
#perception of control and optimism controlling for social desirabilit
ppcor::spcor.test(ydata$tpcoiss, ydata$toptim, ydata$tmarlow)
```

```
##    estimate      p.value statistic   n gp  Method
## 1 0.4830465 5.348979e-26  11.29257 422  1 pearson
```

```
#stress and optimism controlling for social desirability
ppcor::spcor.test(ydata$tpstress, ydata$toptim, ydata$tmarlow)
```

```
##     estimate     p.value statistic   n gp  Method
## 1 -0.4412444 1.73877e-21 -10.06483 422  1 pearson
```

Zero order correlations



```r
#Get zero order correlations as well to fully explore the effect
cor(ydata$tpcoiss, ydata$tpstress)
```

```
## [1] -0.5812413
```

```r
cor(ydata$tpcoiss, ydata$toptim)
```

```
## [1] 0.5161727
```

```r
cor(ydata$tpstress, ydata$toptim)
```

```
## [1] -0.4667559
```

Reporting Results

Partial correlation was used to explore the relationship between perceived control of internal states (as measured by PCOISS) and perceived stress (measured by the Perceived Stress Scale), while controlling for scores on the Marlowe-Crowne Social Desirability Scale. Preliminary analyses were performed to ensure no violation of the assumption of normality, linearity and homoscedasticity. There was a strong, negative partial correlation between perceived control of internal states and perceived stress, controlling for social desirability, ($r = -.53, n = 422, p < .001$). An inspection of the zero-order correlation ($r = -.58$) suggested that controlling for social desirability had very little effect on the strength of the relationship between these two variables.

15mins break ...

Lab Exercise

- Use the 3 models you built in last lab.
- Check the following assumptions:
  - Cook's distance (numerical and plot).
  - Density plots of residuals.
  - Leverage plot.
  - VIF
  - Tolerance
- Create a new model with interaction term for gender and stress.