One Health at UCalgary

UNIVERSITY OF CALGARY

Statistical Modelling with Data

May 23 – June 02, 2023

Instructor: Qing (Leah) Li, Ph.D. Candidate at Cumming School of Medicine

qing.li2@uclagary.ca

Thank you Dr. Thuntida Ngamkham for contributing the contents

Thank you Dr. Qingrun Zhang and Dr. Quan Long for contributing some slides

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression, Forward selection and Backward Elimination
  - Lecture 5: Model selection: Evaluate the reliability of the model chosen
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, equal variance assumptions and normality assumptions.
  - Lecture 7: Multiple regression diagnostics: identify multicollinearity and outliers and data transformation.
- Topic 5: Transfer learning
  - Lecture 8: Deep learning basics
  - Lecture 9: Transfer-learning (Bonus): standing on the shoulders of giants.

Statistical Modelling with Data

- Topic 1: Statistical Modelling
  - Lecture 1: First-order models with quantitative independent variables
- Topic 2: Statistical Modelling with interactions (Assignment 1)
  - Lecture 2: Interaction effects, quantitative and qualitative variables
  - Lecture 3: Interaction effects and second-order models
- Topic 3: Statistical Model selection (Assignment 2)
  - Lecture 4: Model selection: Stepwise regression, Forward selection and Backward Elimination
  - Lecture 5: Model selection: Evaluate the reliability of the model chosen
- Topic 4: Statistical model diagnostics
  - Lecture 6: Multiple regression diagnostics: verify linearity, independence, equal variance assumptions, and normality assumptions.
  - Lecture 7: Multiple regression diagnostics: verify, identify multicollinearity and outliers and data transformation.
- Topic 5: Transfer learning
  - Lecture 8: Deep learning basics
  - Lecture 9: Deep learning advances: Transfer-learning (Bonus).

Statistical Modelling with Data

Learning Outcomes: At the end of the course, participants will be able to

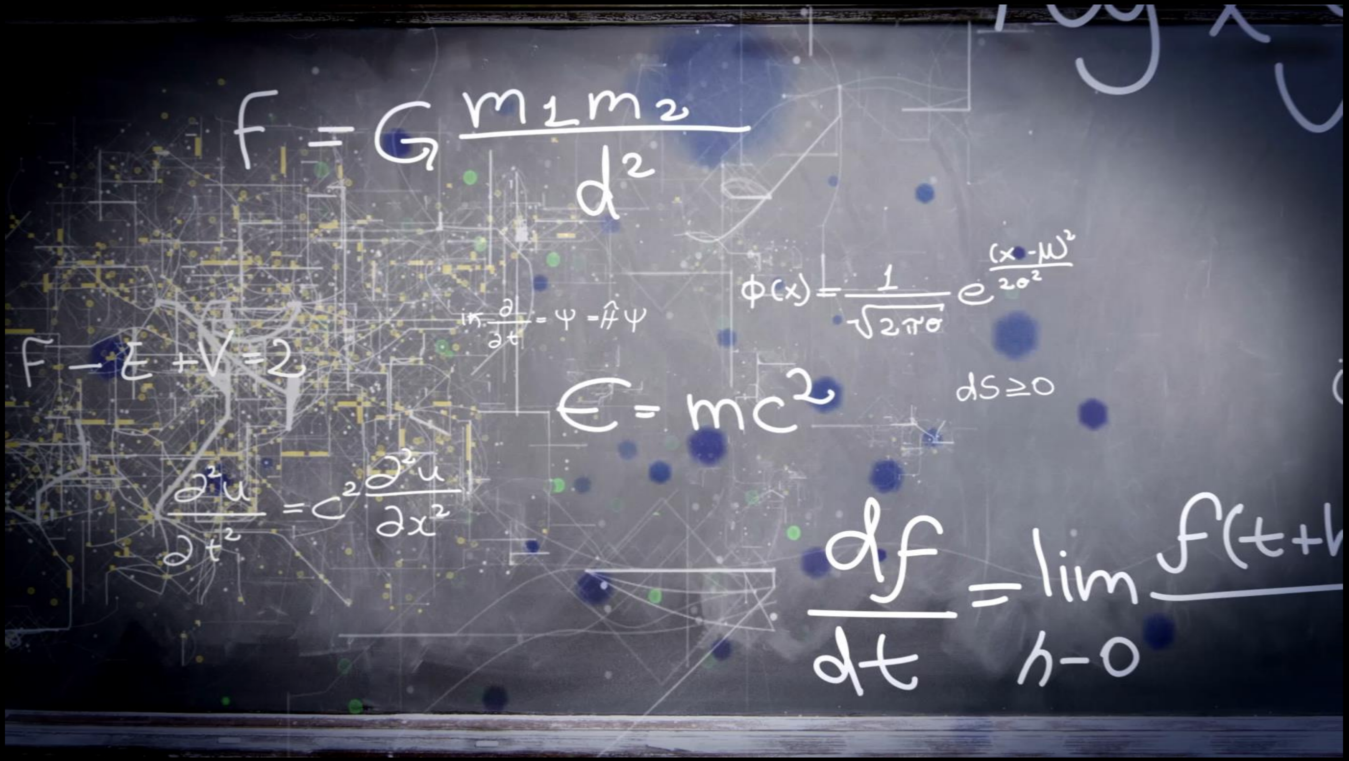
1. Model the multiple linear relationships between a response variable (Y) and all explanatory variables (both categorical and numerical variables) with interaction terms. Interpret model parameter estimates, construct confidence intervals for regression coefficients, evaluate model fits, and visualize correlations between a response variable (Y) and all explanatory variables (X) by graphs (scatter plot, residual plot) to assess model validity.
2. Predict the response variable at a certain level of the explanatory variables once the fit model exists.
3. Implement R-software and analyze statistical results for biomedical and other data.

- **Evaluations**
  1. Assignments will be posted on Slack (our communication tool with students).
  2. Students must attend 70% (6/9) of the sessions in order to receive the certificate and are encouraged to work on the assignments progressively throughout the course as the relevant material is covered.

Statistical Modelling with Data

- Supportive materials
  - Lectures slides (2023)
  - R code scripts (2023)
  - PDF (dated 2022)
  - Two Assignments (dated 2022)
- Slack channels
  - Recoding videos
  - Exercises
  - Course-documents

Lecture 6: Multiple regression diagnostics: verify linearity, independence, and equal variance,


$F = G\frac{m_1 m_2}{d^2}$
$\phi(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$
$F - E + V = 2$
$\epsilon = mc^2$
$dS \geq 0$
$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$
$\frac{df}{dt} = \lim_{h-0} \frac{f(t+h}{}$

Plots

- R: plot()
- Pros:
  - Easy to write and learn by heart
- Cons:
  - Does not incorporate advanced functions such as fitting curves to points
- Multiple figures in one panel
  - par(mfrow=c(nrow, ncol))

- ggplot2 package: ggplot()
- Pros:
  - Multiple choices, high-level functions
- Cons:
  - It takes a while to understand and learn the functions by heart
- Multiple figures in one panel
  - grid.arrange(p1,p2, nrow = 3,top = “Title").

Workflow of regression analysis

1. State a hypothesis
2. Data exploration
3. Build a statistical model
4. Model Diagnostics:
    1. Residual VS fitted plot, Normal QQ plot of residuals, added variables plots, influence plot (residual vs. leverage)
5. Fix the biggest problem and go back to 3
6. Compare alternative models with reduced model
7. Interpret the regression coefficients are the effect large and what they mean in your research context.

Workflow of regression analysis (modified by Leah)

1. State a hypothesis
2. Data exploration
3. Build statistical models (different approaches)
4. Diagnose if the optimal model from different approaches violet assumptions or not. If not, these model are validated.
5. If not, fix the problem and conduct model diagnostic until identify a model satisfy all assumptions.
6. Interpret the regression coefficients are the effect large and what they mean in your research context.

What are model diagnostics?

- Most statistical tests rely upon certain assumptions about the variables used in the analysis. **When these assumptions are not met the results may not be trustworthy**. The assumptions and conditions for the multiple regression model sound nearly the same as for simple linear regression, but with more variables in the model.
- Even if you have a solid theoretical basis for your assumptions, they may still turn out to be incorrect.
- Model diagnostics are visual and numeric guides to:
  - Help decide whether the assumptions taken hold up in all samples;
  - Provide an indication of why an assumption is violated.
- Model diagnostics are not:
  - An indication that the assumption are correct
  - An indication that the model is right or wrong. We hope to find our optimal model does not violate any of these assumptions.
- Assumptions must be based on theory. No test or visual diagnostic plot can tell you which assumption to choose, only whether they hold up reasonably well in your studies samples or not.

Assumptions

response, coefficients of correlation, predictors

$$g(y_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \mathrm{g}(y_i) = y_i$$

1. Linearity Assumption

   Relationships between all predictors and the response are linear

2. Independence Assumption

   Independence of observations

3. Equal Variance Assumption

   Error term (Residuals) has equal variance given any values of independent variables

4. Normality Assumption

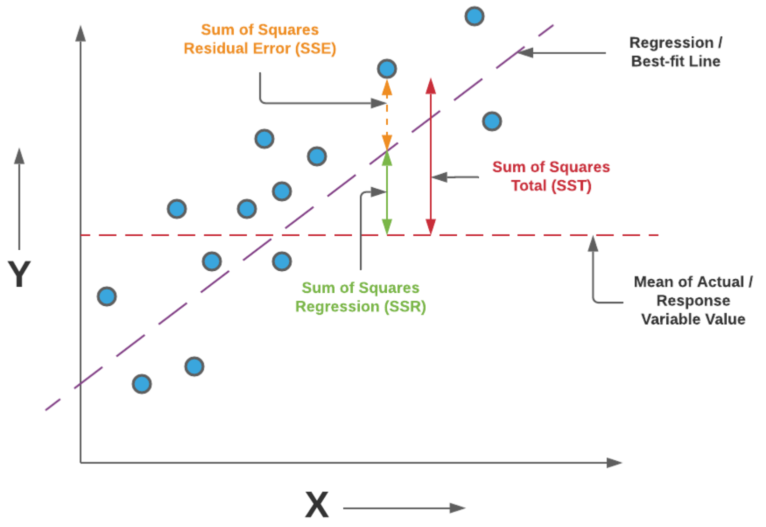
   Error term (Residuals) is normally distributed

5. Outlier

   Error terms (Residuals) has expected value of zero given the values of independent va

6. Multicollinearity

   No perfect collinearity and non-zero variance of independent variables

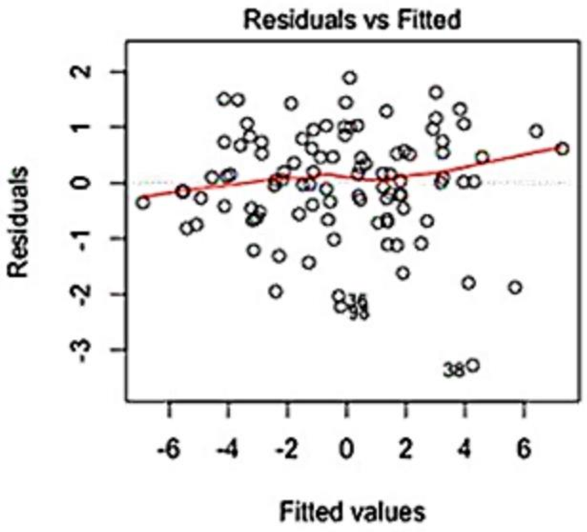

Sum of Squares
Residual Error (SSE)
Regression / Best-fit Line
Sum of Squares Total (SST)
Y
Sum of Squares Regression (SSR)
Mean of Actual / Response Variable Value
X

1. Linearity Assumption

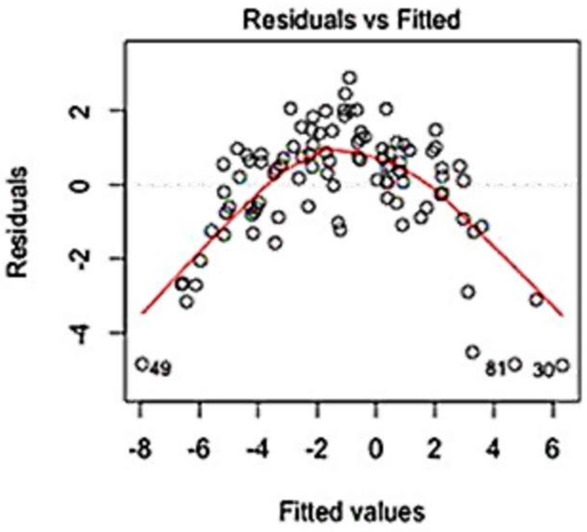
The linear regression model assumes that there is a straight-line relationship between the predictors and the response. If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspicious. In addition, the prediction accuracy of the model can be significantly reduced.

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as log(X), Sqrt(X), and X^2 in the regression model.

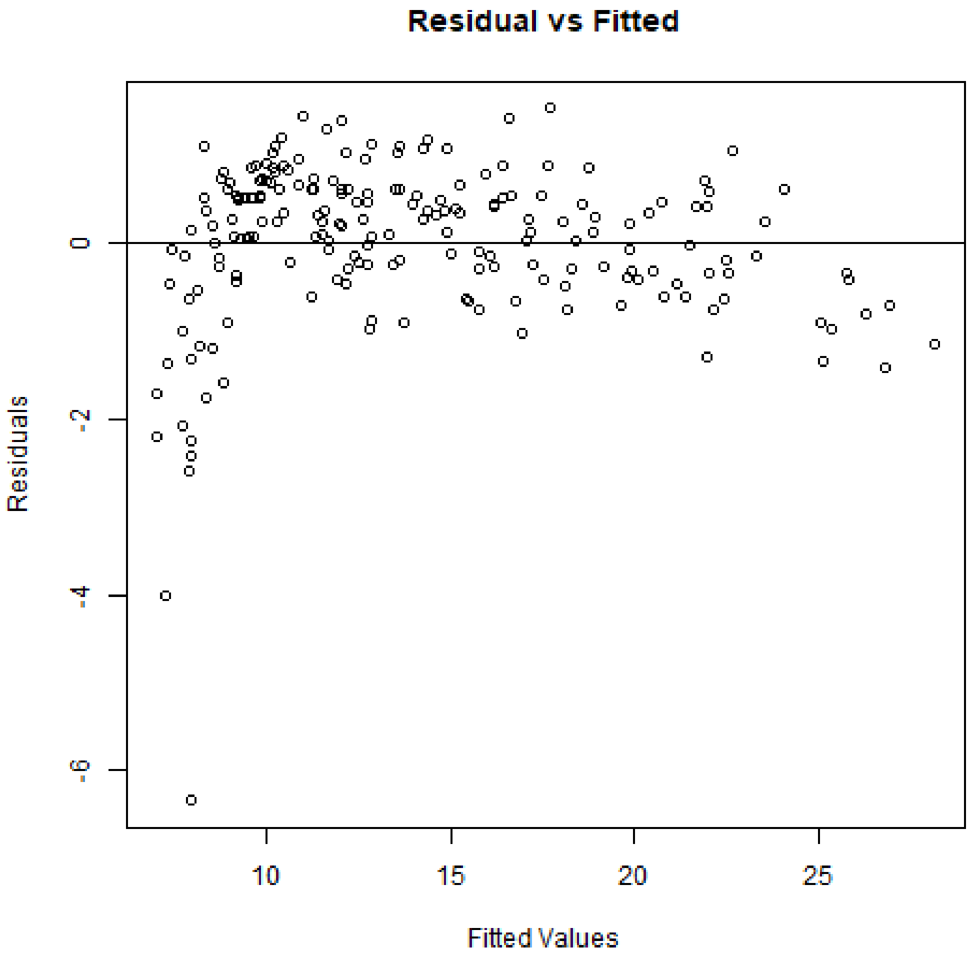
No Pattern Evident



Non-Linearity Evident



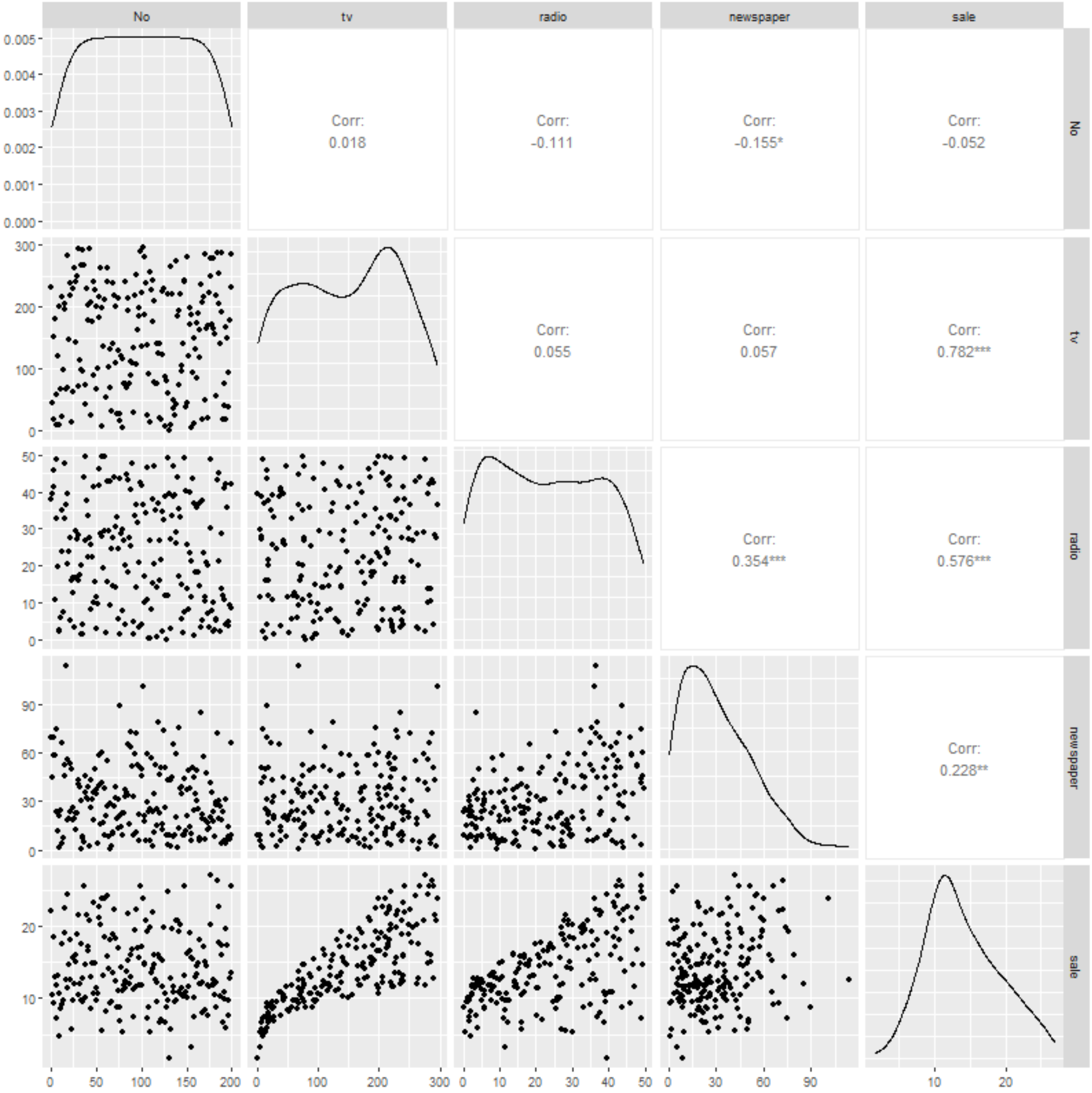
Advertising data

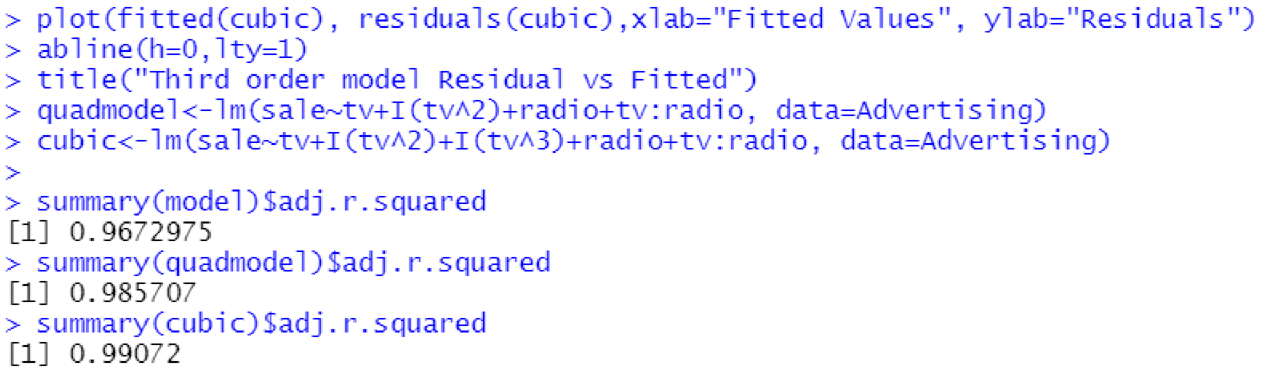


model<-lm(sale~tv+radio+tv:radio, data=Advertising)

There appears to be a little pattern in the residuals, suggesting that the quadratic term or logarithmic might improve the fit to the data.

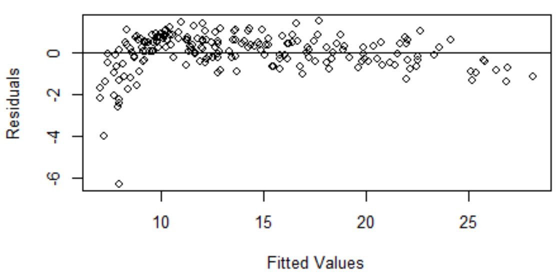
Advertising data: Pairwise correlation



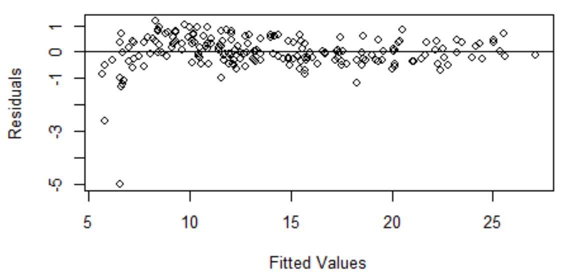


```
> plot(fitted(cubic), residuals(cubic),xlab="Fitted Values", ylab="Residuals")
> abline(h=0,lty=1)
> title("Third order model Residual vs Fitted")
> quadmodel<-lm(sale~tv+I(tv^2)+radio+tv:radio, data=Advertising)
> cubic<-lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv:radio, data=Advertising)
>
> summary(model)$adj.r.squared
[1] 0.9672975
> summary(quadmodel)$adj.r.squared
[1] 0.985707
> summary(cubic)$adj.r.squared
[1] 0.99072
```

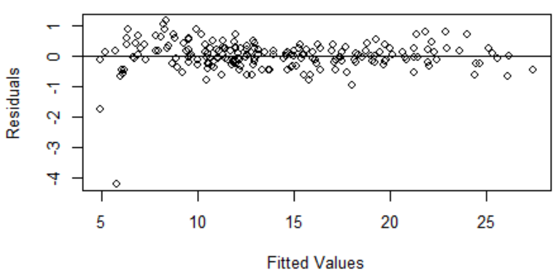
Interaction model Residual vs Fitted



Second order model Residual vs Fitted



Third order model Residual vs Fitted



Conclusion: residuals from third order model do not have an obvious pattern with fitted values.

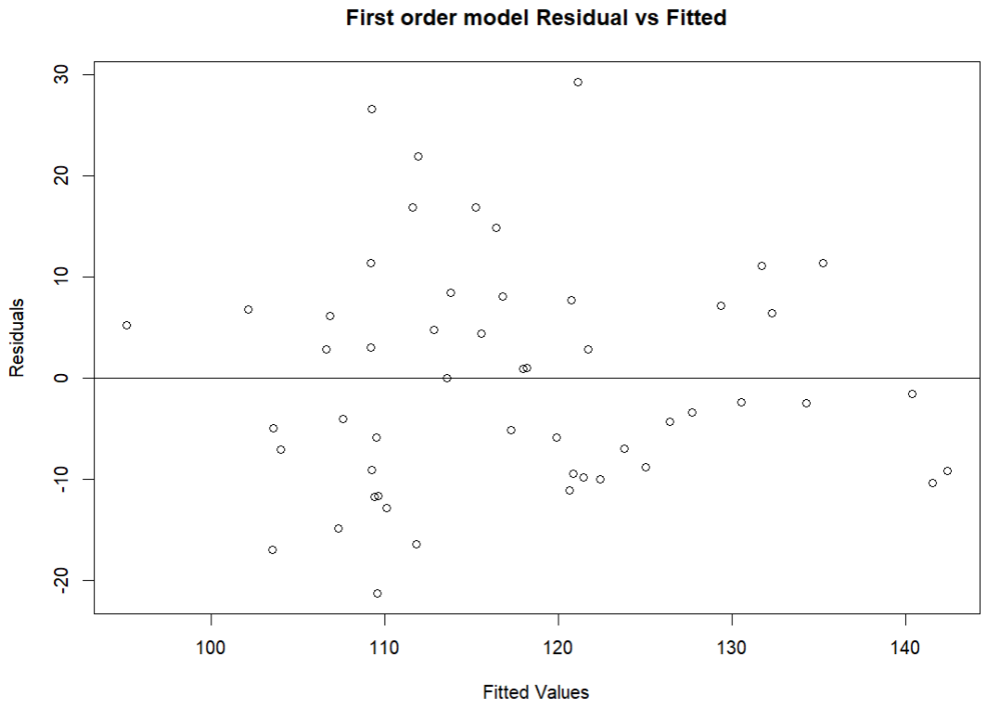
In class Practice Problem 16

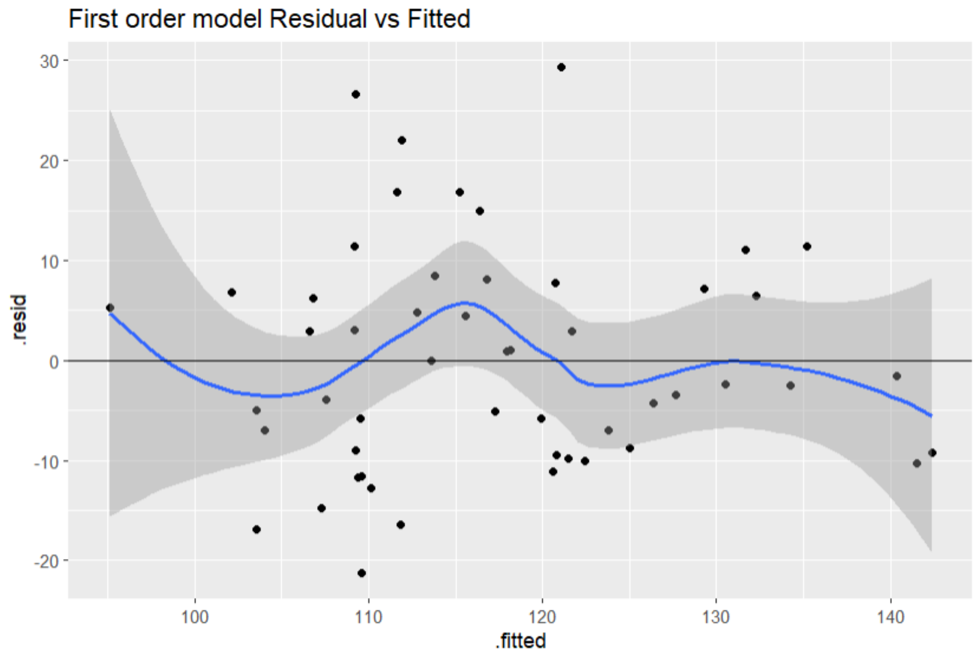
From the clerical staff work hours, use residual plots to conduct a residual analysis of the data.
Begin with the model: Y ~ X2 + X4 + X5

1. Check whether this model meets the linearity assumption
2. If it doesn't (it doesn't), use ggpairs() to identify potential terms that might be transformed in a higher order model.
3. Fit that higher order model and evaluate.

☞ Lecture_6.R

In class Practice Problem 16 Answers





Pattern is not quite obvious with plot, but observable. Maybe better with ggplot() as it can fit a line to data

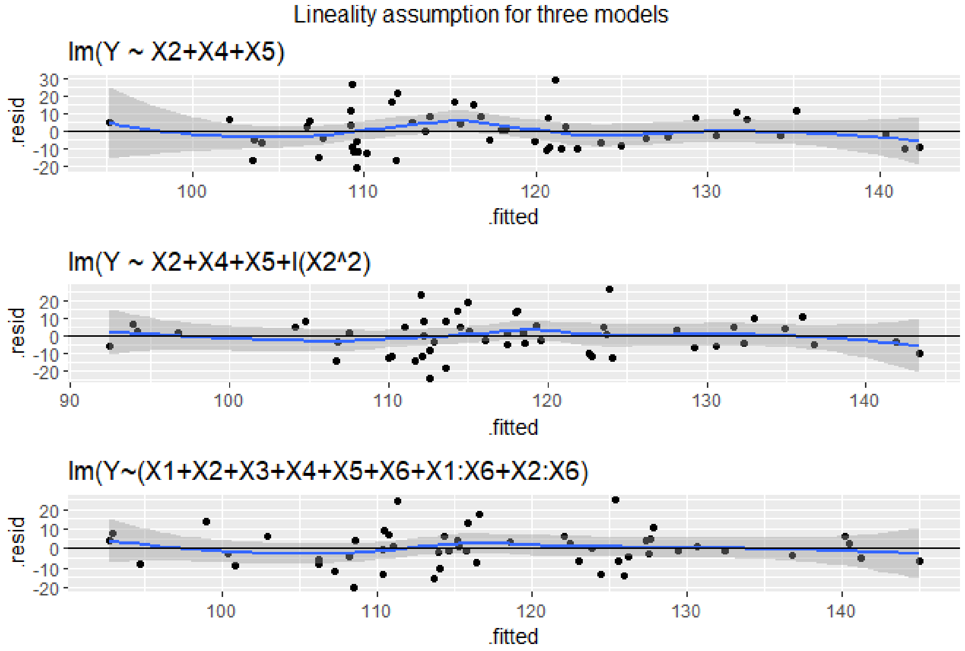
Based on these two figures, we observed patterns between residuals of the model and response.

>>This model does not satisfy the linearity assumption.

Lecture_6.R

In class Practice Problem 16 Answers

Optimal models from the problem 15:
optimal_model_approach1=lm(Y ~ X2+X4+X5+I(X2^2), data=workhours)
optimal_model_approach2=lm(Y ~ X2+X4+X5+I(X2^2), data=workhours)
optimal_model_approach3 <-lm(Y~(X1+X2+X3+X4+X5+X6+X1:X6+X2:X6), data=workhours)



The bottom two models appear to meet linearity assumption.

Lecture_6.R

2. Independence Assumption

An important assumption of the linear regression model is that the error terms, $\epsilon 1, \epsilon 2, \epsilon 3, ..., \epsilon n$ are uncorrelated (must be mutually independent). What does this mean? For instance, if the errors are uncorrelated, then the fact that ϵi is positive provides little or no information about the sign of $\epsilon i+1$.
The assumption of independent errors is violated when successive errors are correlated. This typically occurs when the data for both dependent and independent variables are observed sequentially over a period of time- called **time-series data**

We can check displays of the regression residuals for evidence of patterns, trends or clumping, any of which would suggest a failure of independence. In the special case when response *Y* is related to time (time series data), a common violation of the Independence Assumption is for the errors to be correlated. This violation can be checked by plotting the residuals against the order of occurrence (time plot of the residuals and looking for pattern).

In the Advertising example, the subjects were not related to time, so we can pretty sure that their measurements are independent.

Independence Assumption

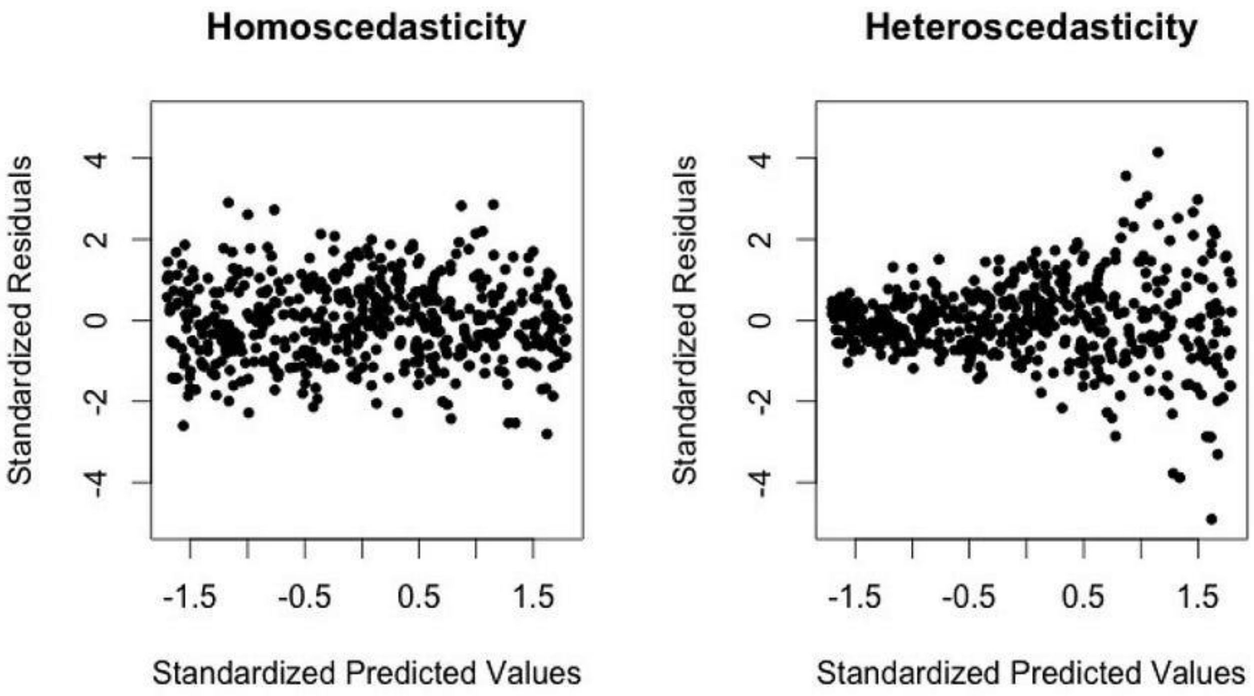
- Lack of independence can be caused by:
  - Serial correlation (whether this month's temperature is correlated with the one in the last month's)
  - Spatial association
  - Cluster effect

- Make plots of residuals vs relevant explanatory variables and look for pattern
  - Residuals vs time (or observation number)
  - Residuals vs spatial variable
  - Residuals vs group (prefer blocking)

3. Equal Variance Assumption

Another important assumption of the linear regression model is that the error terms have a constant variance (**homoscedasticity**), *Var*(*ϵi*) = *σ*2. Unfortunately, it is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response.

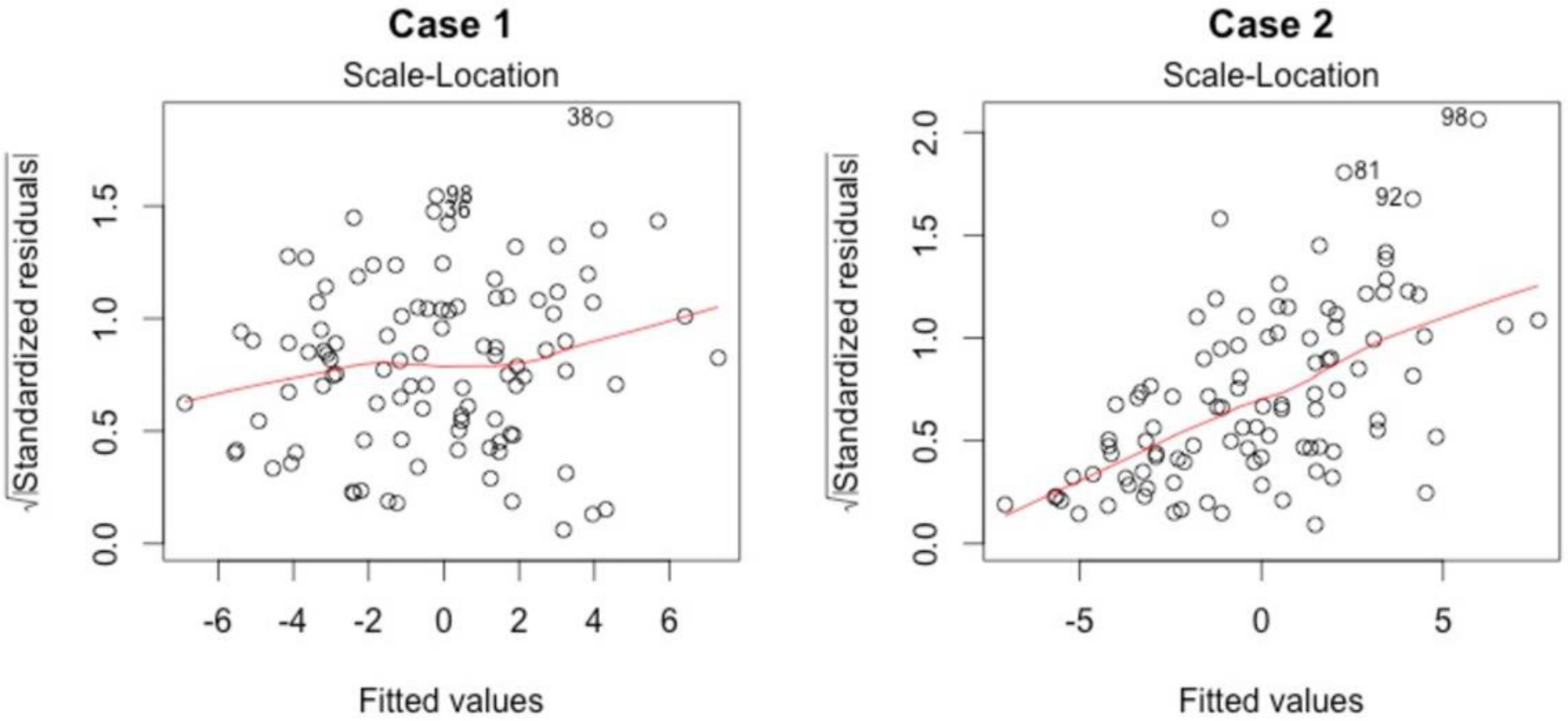
One can identify non-constant variances in the errors, or **heteroscedasticity**



Heteroscedasticity

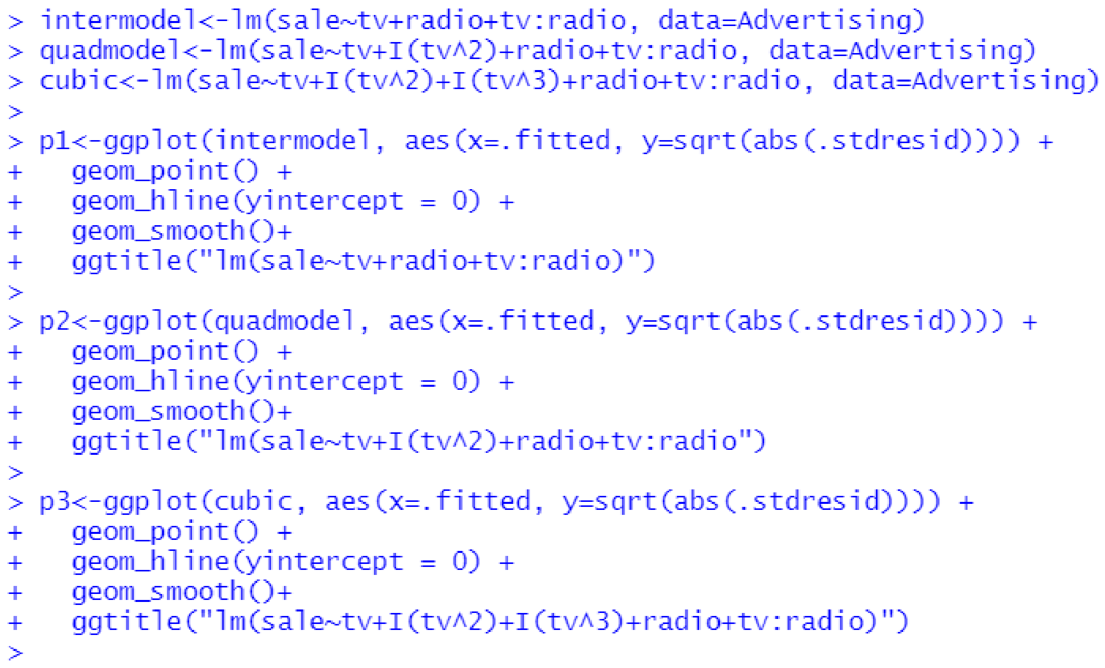
- Result of Heteroscedasticity will be an inefficient estimator
  - Large SE of the estimator
- Reasons for heteroscedasticity
  - The sample size different for each Y
  - Variance or standard error is a constant percentage of the y-value
- How to identify if a model have heteroscedasticity or not?
  - A scale-location plot between fitted value and standardized residuals can also be checked for heteroscedasticity. It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. You can check the assumption of equal variance (homoscedasticity).

The scale location plot

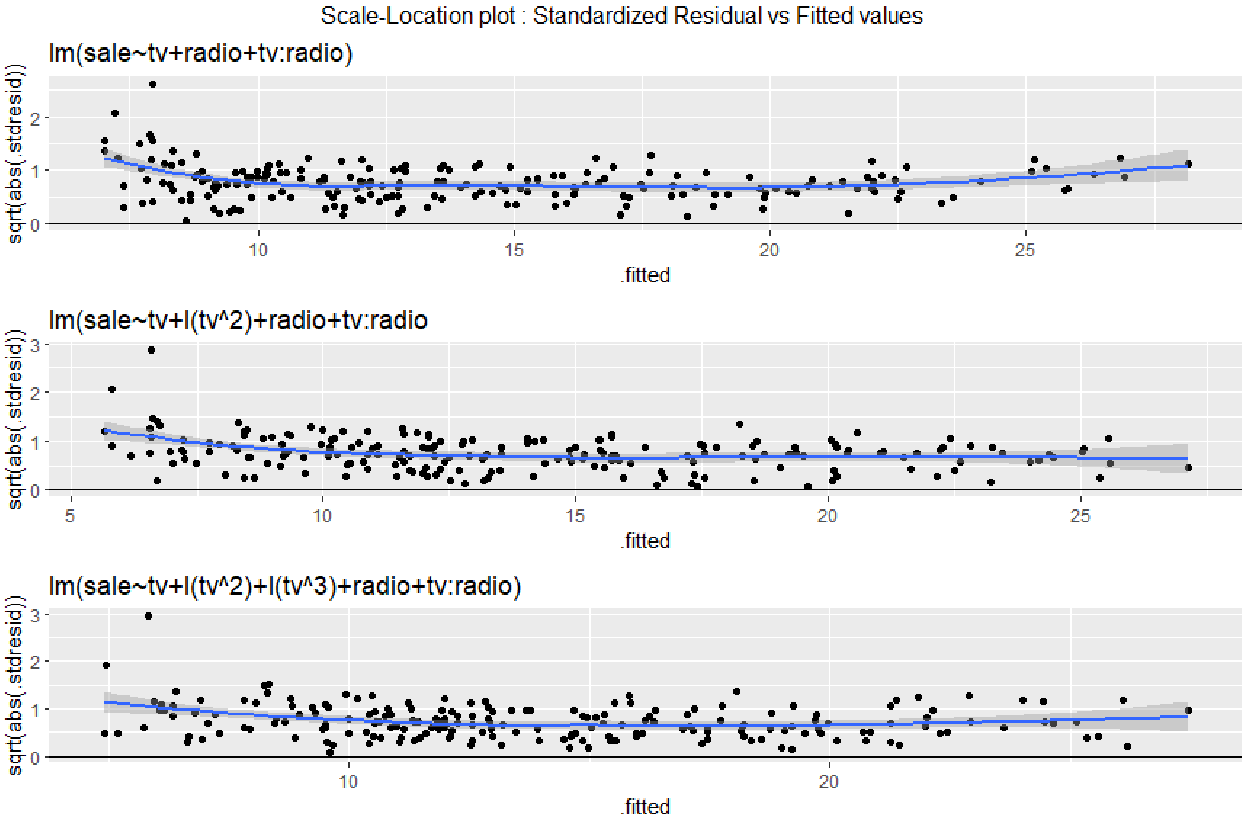


It’s good if you see a horizontal line with equally (randomly) spread points. From the figure above in Case 1, the residuals appear randomly spread. Whereas, in Case 2, the residuals begin to spread wider along the x-axis as it passes around 5. Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2.

Advertising data Heteroscedasticity



```
> intermodel<-lm(sale~tv+radio+tv:radio, data=Advertising)
> quadmodel<-lm(sale~tv+I(tv^2)+radio+tv:radio, data=Advertising)
> cubic<-lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv:radio, data=Advertising)
>
> p1<-ggplot(intermodel, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
+   geom_point() +
+   geom_hline(yintercept = 0) +
+   geom_smooth()+
+   ggtitle("lm(sale~tv+radio+tv:radio)")
>
> p2<-ggplot(quadmodel, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
+   geom_point() +
+   geom_hline(yintercept = 0) +
+   geom_smooth()+
+   ggtitle("lm(sale~tv+I(tv^2)+radio+tv:radio")
>
> p3<-ggplot(cubic, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
+   geom_point() +
+   geom_hline(yintercept = 0) +
+   geom_smooth()+
+   ggtitle("lm(sale~tv+I(tv^2)+I(tv^3)+radio+tv:radio)")
>
```



From the Advertising example, the output displays the residual plot and Scale-Location plot that result from the cubic model. In our case, the residuals tend to form a horizontal band-indicates that the plot does not provide evidence to suggest that heteroscedasticity exists.

The Breusch-Pagan Test

A more formal, mathematical way of detecting heteroscedasticity is what is known as **the Breusch-Pagan test**. It involves using a variance function and using a $\chi^2 test$ to test

$$H_0 : \text{heteroscedasticity is not present (homoscedasticity)}$$

$$H_a : \text{heteroscedasticity is present}$$

$$or$$

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$$

$$H_0 : \text{at least } \sigma_i^2 \text{ is different from the others } i = 1, 2, \ldots, n$$

$$\chi^2 = nR^2 \sim \chi^2_{p-1}$$

library(lmtest)

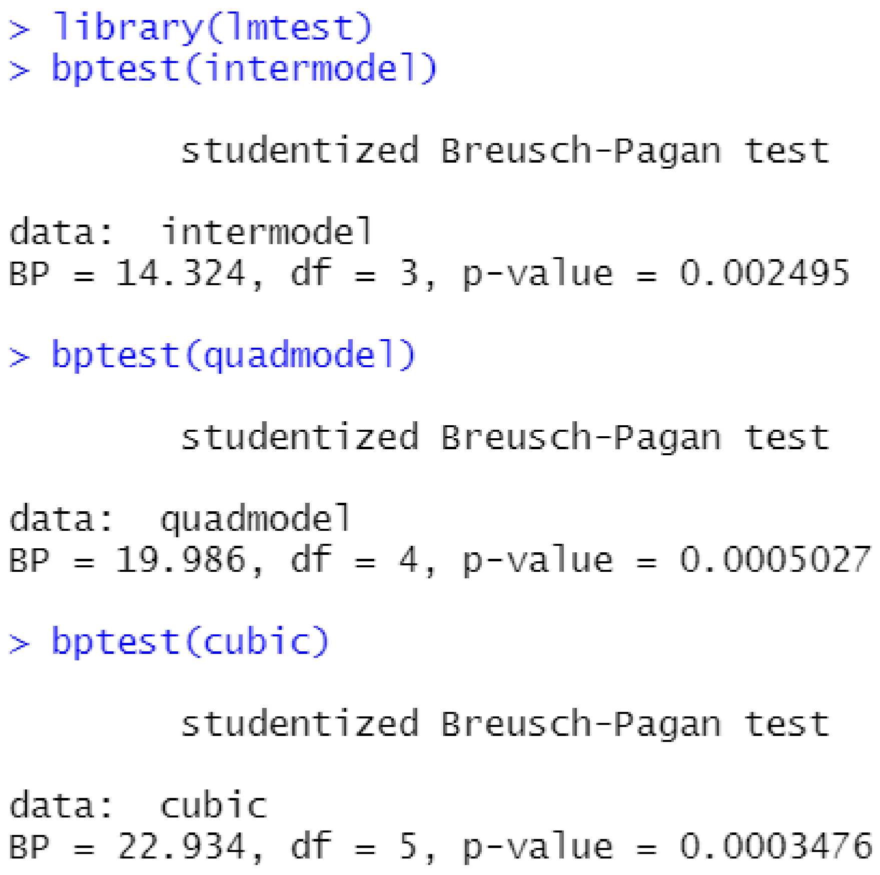
$$where$$

$$n = \text{sample size}$$

$$R^2 = \text{coefficient determination}$$

$$p = \text{ number of regression coefficients}$$

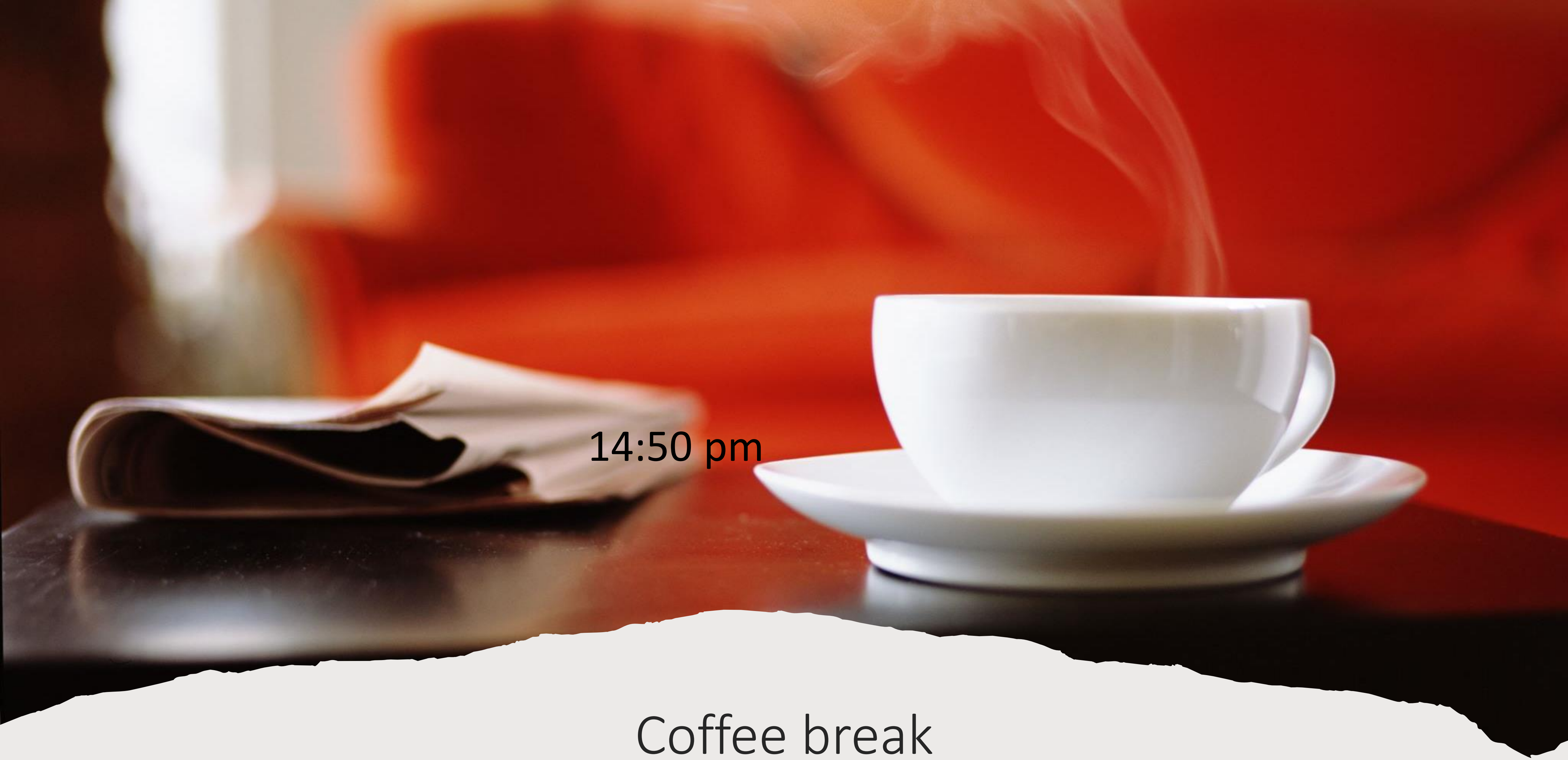
Advertising data



```
> library(lmtest)
> bptest(intermodel)

	studentized Breusch-Pagan test

data:  intermodel
BP = 14.324, df = 3, p-value = 0.002495

> bptest(quadmodel)

	studentized Breusch-Pagan test

data:  quadmodel
BP = 19.986, df = 4, p-value = 0.0005027

> bptest(cubic)

	studentized Breusch-Pagan test

data:  cubic
BP = 22.934, df = 5, p-value = 0.0003476
```

From the Advertising example, the Breusch-Pagan test p-value are Signiant ($p<0.05$). We do reject the null hypothesis, which is no presence of Heteroscedasticity.

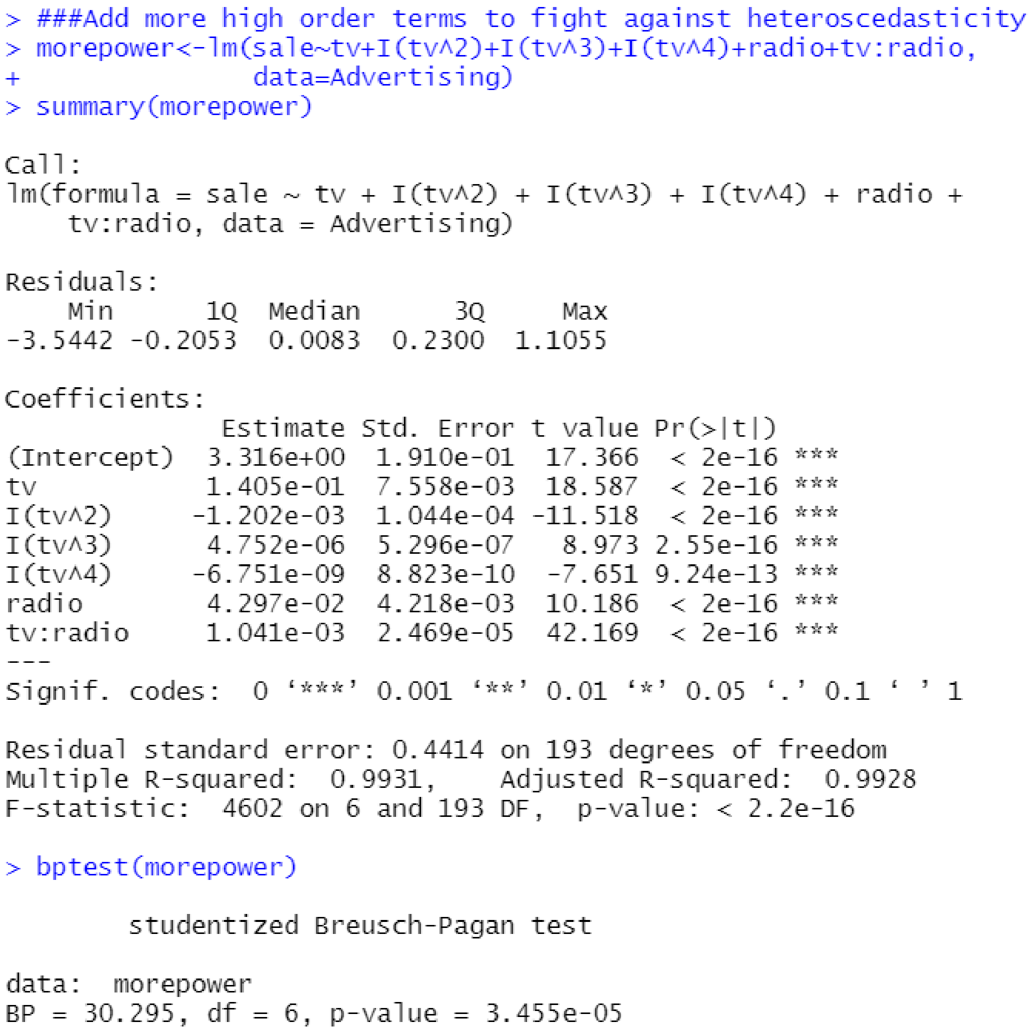
What can we do now?



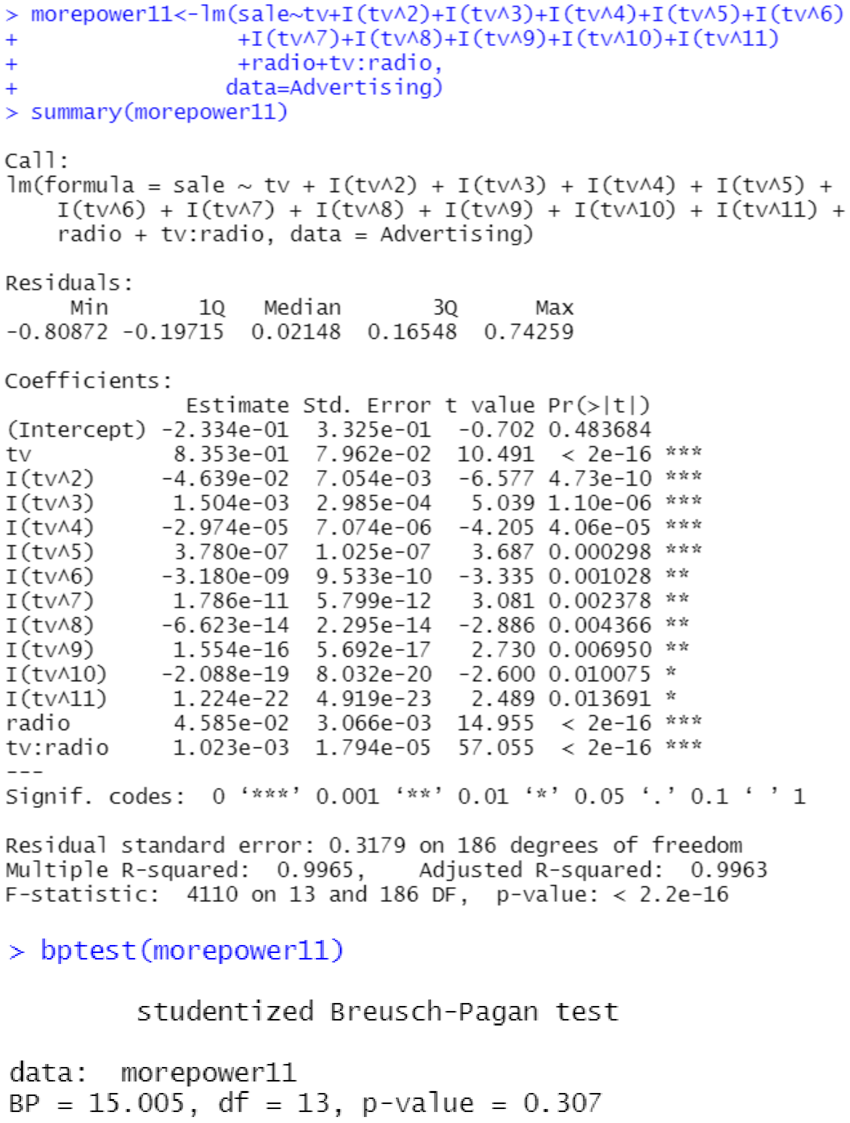
14:50 pm

Coffee break

High orders to eliminate Heteroscedasticity



```
> ###Add more high order terms to fight against heteroscedasticity
> morepower<-lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+radio+tv:radio,
+               data=Advertising)
> summary(morepower)

Call:
lm(formula = sale ~ tv + I(tv^2) + I(tv^3) + I(tv^4) + radio +
    tv:radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5442 -0.2053  0.0083  0.2300  1.1055

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.316e+00  1.910e-01  17.366  < 2e-16 ***
tv           1.405e-01  7.558e-03  18.587  < 2e-16 ***
I(tv^2)     -1.202e-03  1.044e-04 -11.518  < 2e-16 ***
I(tv^3)      4.752e-06  5.296e-07   8.973 2.55e-16 ***
I(tv^4)     -6.751e-09  8.823e-10  -7.651 9.24e-13 ***
radio        4.297e-02  4.218e-03  10.186  < 2e-16 ***
tv:radio     1.041e-03  2.469e-05  42.169  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4414 on 193 degrees of freedom
Multiple R-squared:  0.9931,	Adjusted R-squared:  0.9928
F-statistic:  4602 on 6 and 193 DF,  p-value: < 2.2e-16

> bptest(morepower)

	studentized Breusch-Pagan test

data:  morepower
BP = 30.295, df = 6, p-value = 3.455e-05
```



```
> morepower11<-lm(sale~tv+I(tv^2)+I(tv^3)+I(tv^4)+I(tv^5)+I(tv^6)
+                 +I(tv^7)+I(tv^8)+I(tv^9)+I(tv^10)+I(tv^11)
+                 +radio+tv:radio,
+                data=Advertising)
> summary(morepower11)

Call:
lm(formula = sale ~ tv + I(tv^2) + I(tv^3) + I(tv^4) + I(tv^5) +
    I(tv^6) + I(tv^7) + I(tv^8) + I(tv^9) + I(tv^10) + I(tv^11) +
    radio + tv:radio, data = Advertising)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80872 -0.19715  0.02148  0.16548  0.74259

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.334e-01  3.325e-01  -0.702 0.483684
tv           8.353e-01  7.962e-02  10.491  < 2e-16 ***
I(tv^2)     -4.639e-02  7.054e-03  -6.577 4.73e-10 ***
I(tv^3)      1.504e-03  2.985e-04   5.039 1.10e-06 ***
I(tv^4)     -2.974e-05  7.074e-06  -4.205 4.06e-05 ***
I(tv^5)      3.780e-07  1.025e-07   3.687 0.000298 ***
I(tv^6)     -3.180e-09  9.533e-10  -3.335 0.001028 **
I(tv^7)      1.786e-11  5.799e-12   3.081 0.002378 **
I(tv^8)     -6.623e-14  2.295e-14  -2.886 0.004366 **
I(tv^9)      1.554e-16  5.692e-17   2.730 0.006950 **
I(tv^10)    -2.088e-19  8.032e-20  -2.600 0.010075 *
I(tv^11)     1.224e-22  4.919e-23   2.489 0.013691 *
radio        4.585e-02  3.066e-03  14.955  < 2e-16 ***
tv:radio     1.023e-03  1.794e-05  57.055  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3179 on 186 degrees of freedom
Multiple R-squared:  0.9965,	Adjusted R-squared:  0.9963
F-statistic:  4110 on 13 and 186 DF,  p-value: < 2.2e-16

> bptest(morepower11)

	studentized Breusch-Pagan test

data:  morepower11
BP = 15.005, df = 13, p-value = 0.307
```

How to deal with heteroscedasticity

- If you know how the variance changes with each sample
  - Weighted Least Square Regression
- If the variance has unknown dependence on y_i
  - Other regression methods (GMM, generalized method of moments estimation)
- If you know the general trend of variance change with the predictor variable, then you can transform the data
  - Log-transformation
  - Box-Cox transformations
  - These will be studied later…

In class Practice Problem 17

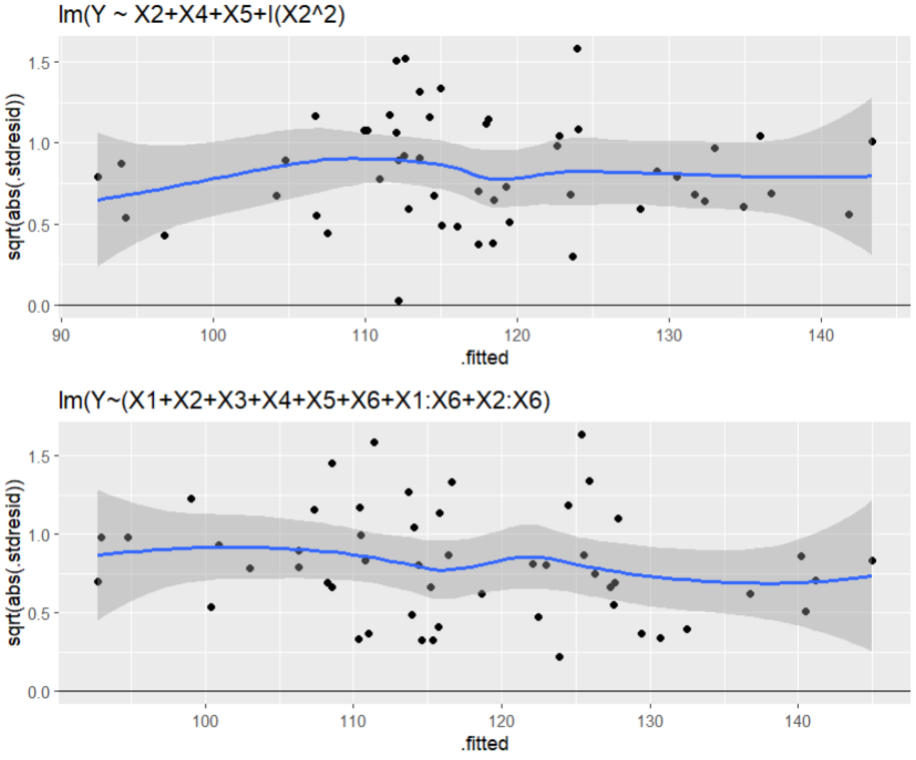
Use the CLERICAL.CSV data.
BEGIN with the best model from PROBLEM 16
Y ~ X2 + I(X2^2) + X4 + X5
Does this model meet the equal-variance assumption?

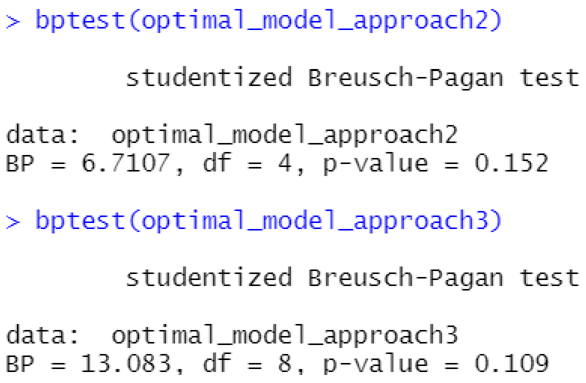
1. Examine residual plot and scale-location plot
2. Conduct the Breusch-Pagan test.

☞ Lecture_6.R

In class Practice Problem 17 Answers

CLERICAL data: Scale-Location plot : Standardized Residual vs Fitted values





```
> bptest(optimal_model_approach2)

	studentized Breusch-Pagan test

data:  optimal_model_approach2
BP = 6.7107, df = 4, p-value = 0.152

> bptest(optimal_model_approach3)

	studentized Breusch-Pagan test

data:  optimal_model_approach3
BP = 13.083, df = 8, p-value = 0.109
```

P value 0.152 and 0.109, > 0.05, cannot reject H0.
So we accept H0, which means the error term of model 2 and 3 does not have heteroscedasticity

Lecture_6.R

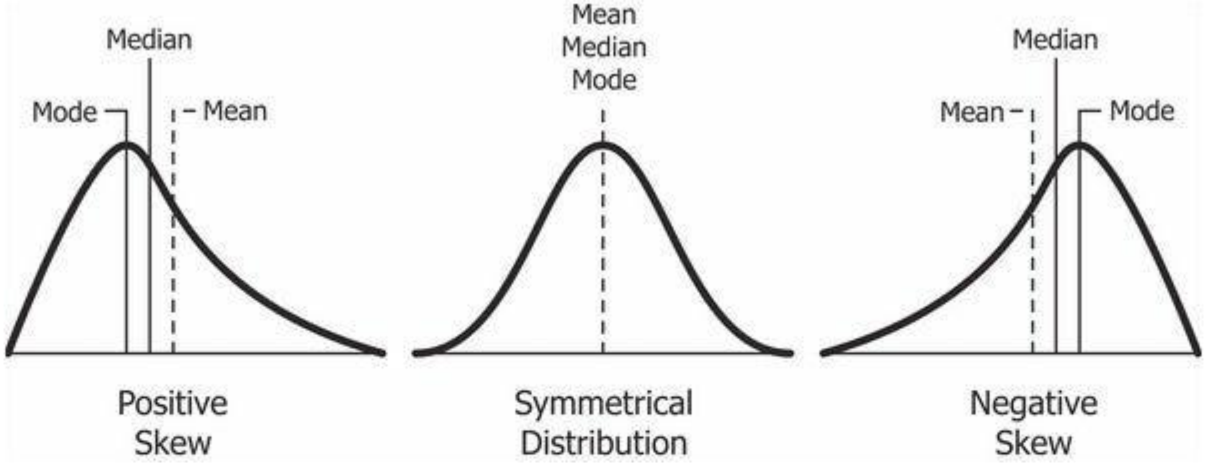
4. Normality Assumption

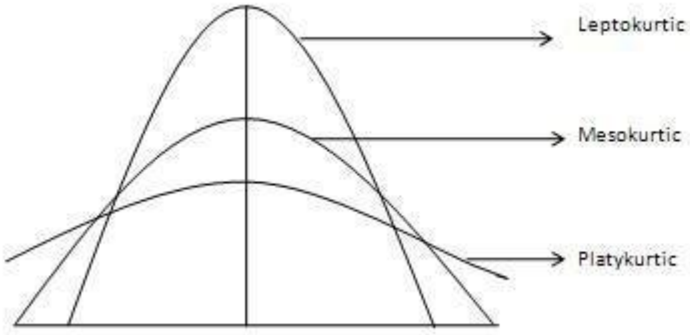
- The multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed.
- This assumption may be checked by looking at a histogram, a normal probability plot or a Q-Q-Plot. If the distribution is normal, the points on such a plot (Probability Plot or Q-Q-Plot) should fall close to the diagonal reference line. A bow-shaped pattern of deviations from the diagonal indicates that the residuals have excessive skewness. An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis, i.e., there are either too many or two few large errors in both directions. Sometimes the problem is revealed to be that there are a few data points on one or both ends that deviate significantly from the reference line ("outliers"), in which case they should get close attention.
- There are also a variety of statistical tests for normality, including the Kolmogorov-Smirnov test and the Shapiro-Wilk test.

$$\begin{aligned} H_0 &: \text{the sample data are significantly normally distributed} \\ H_a &: \text{the sample data are not significantly normally distributed} \end{aligned}$$

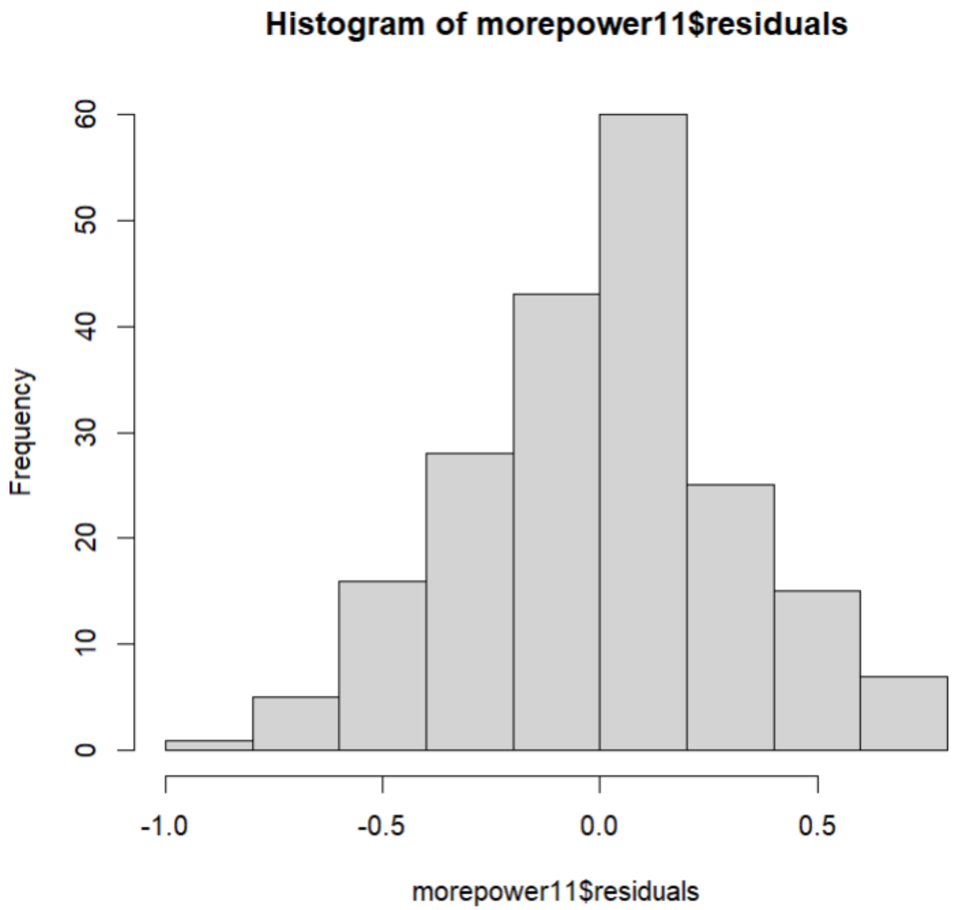
Skewness & Kurtosis

- Skewness is a measure of symmetry, or more precisely, the lack of symmetry
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
    - Data sets with medium kurtosis (Mesokurtic, k=3) has kurtosis statistic similar to that of the normal distribution.
    - Data sets with high kurtosis (Leptokurtic, k<3) tend to have heavy tails, or outliers.
    - Data sets with low kurtosis ((Platykurtic, k>3)) tend to have light tails, or lack of outliers.
- The histogram is an effective graphical technique for showing both the skewness and kurtosis of data set.

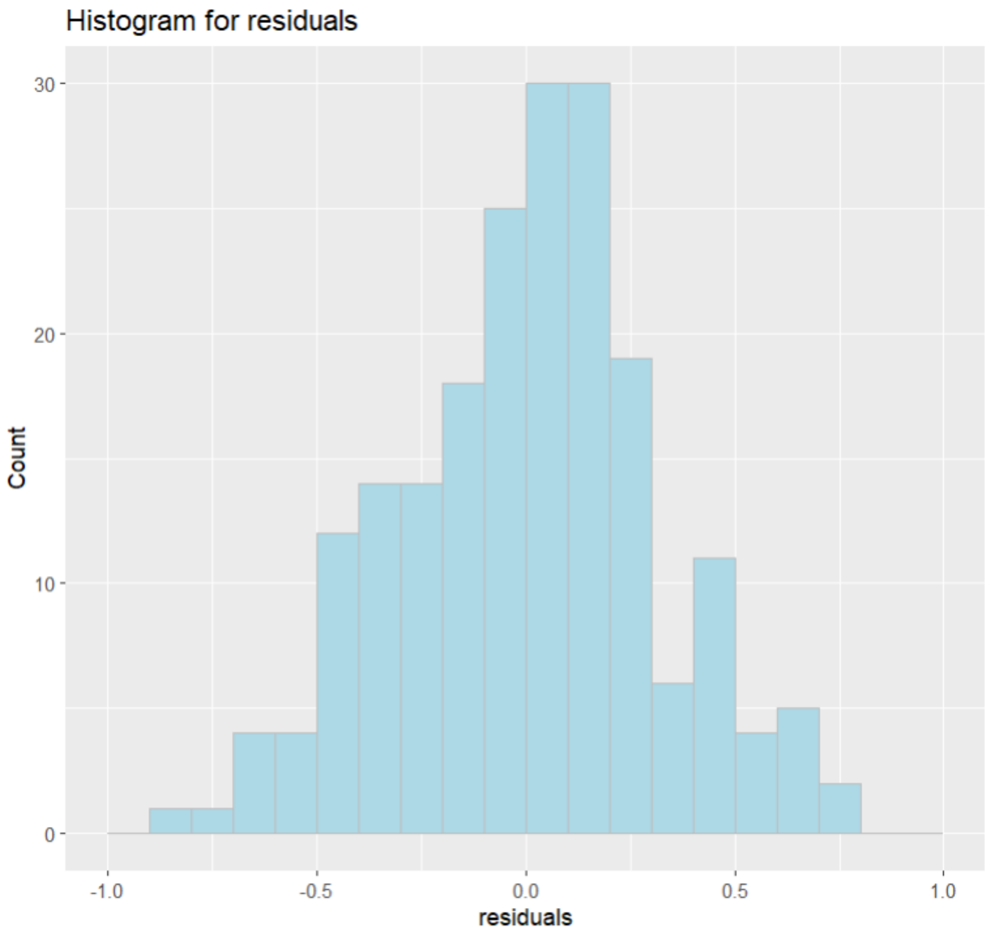




Advertising data Histogram

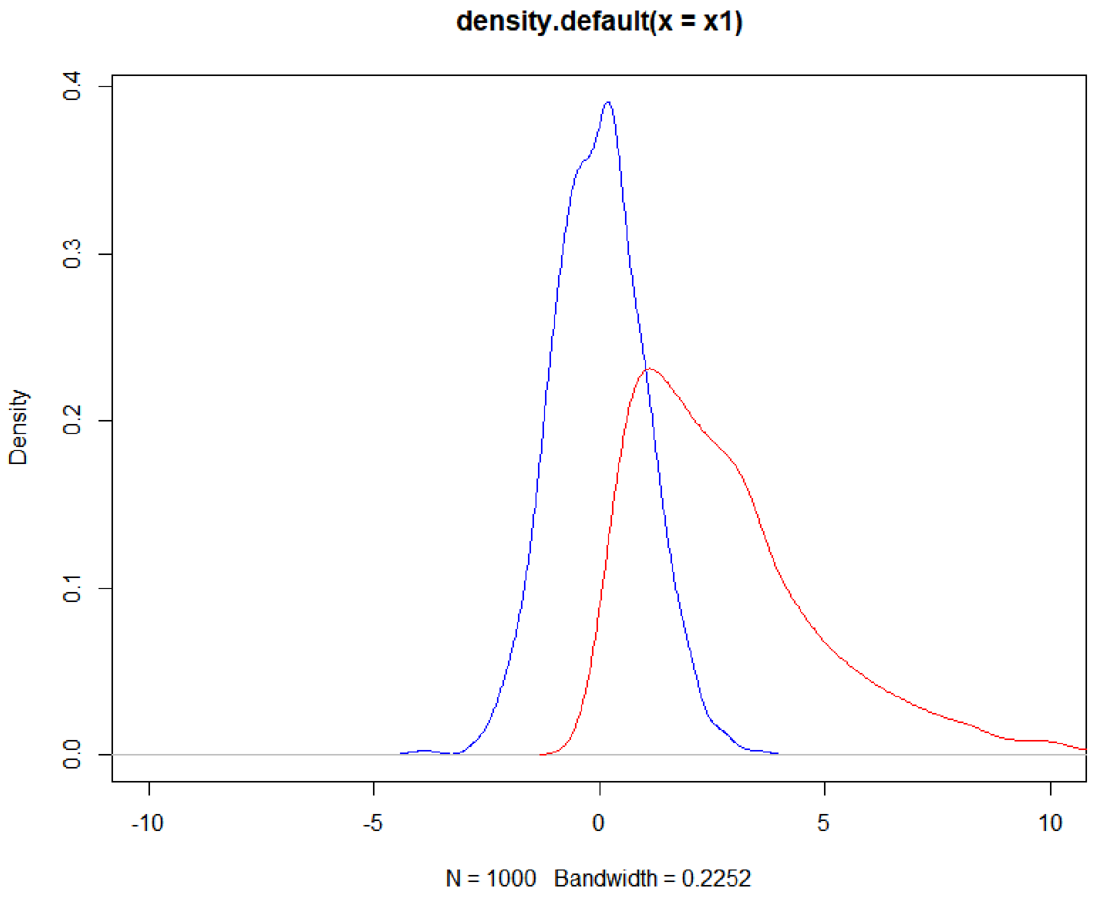


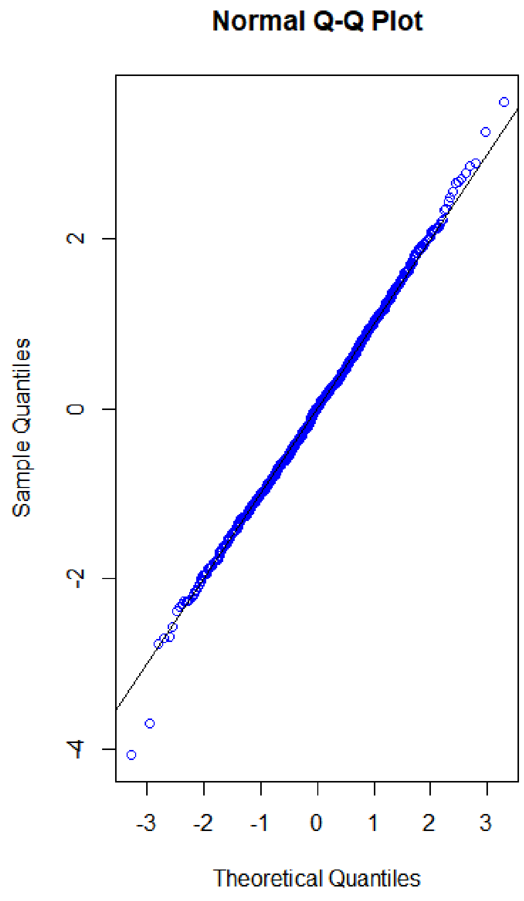
hist(morepower11$residuals)

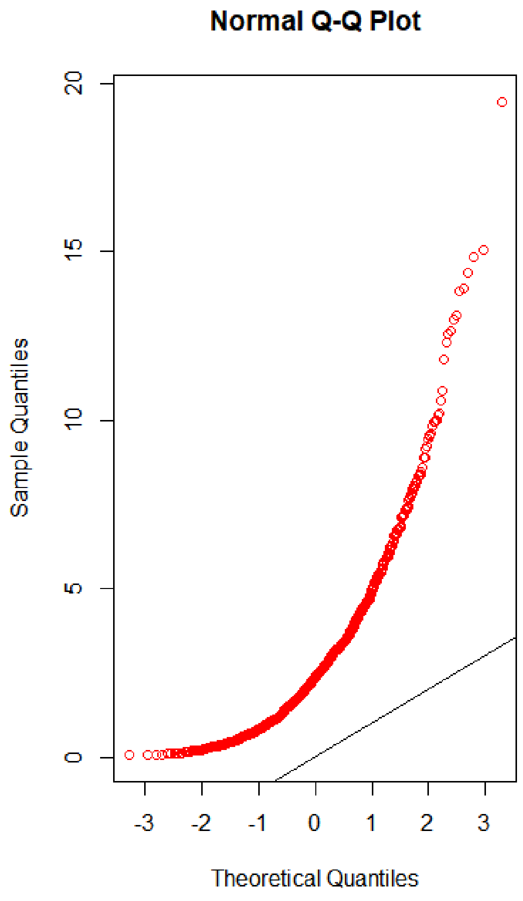


ggplot(data=Advertising, aes(residuals(morepower11))) + geom_histogram(breaks = seq(-1,1,by=0.1), col="grey", fill="lightblue") + labs(title="Histogram for residuals") +labs(x="residuals", y="Count")

Normal Probability Plot



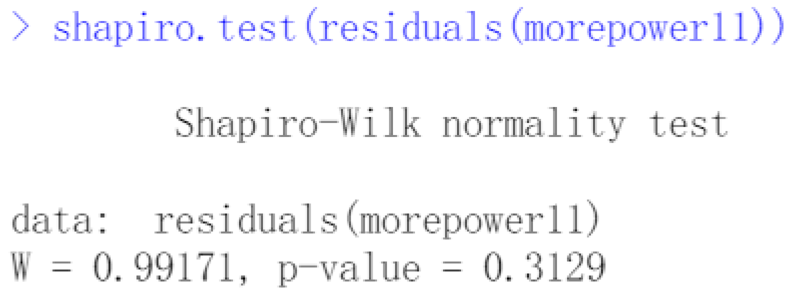




Shapiro Wilk Test

H_0 : the sample data are significantly normally distributed

H_a : the sample data are not significantly normally distributed



```
> shapiro.test(residuals(morepower11))

        Shapiro-Wilk normality test

data:  residuals(morepower11)
W = 0.99171, p-value = 0.3129
```

Kolmogorov-Smirnov test

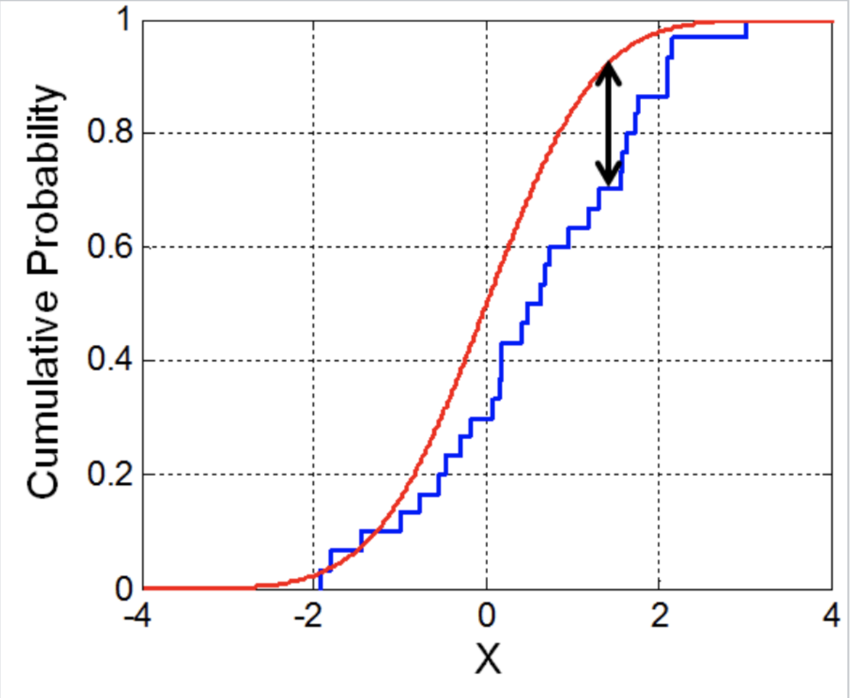
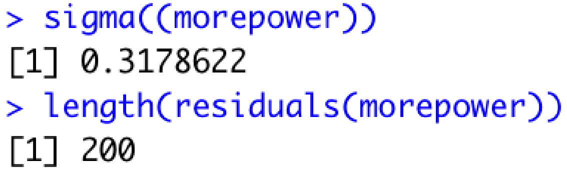
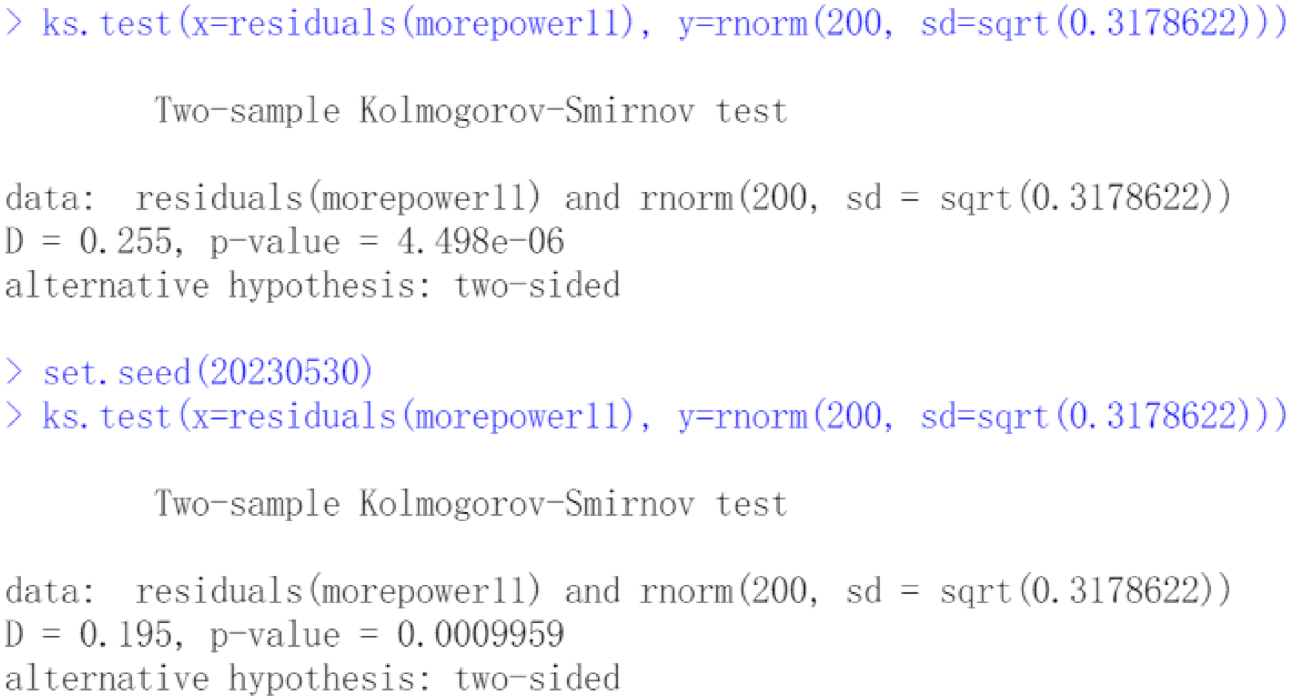


Illustration of the Kolmogorov–Smirnov statistic. The red line is a model CDF, the blue line is an empirical CDF, and the black arrow is the K–S statistic.



```
> sigma((morepower))
[1] 0.3178622
> length(residuals(morepower))
[1] 200
```



```
> ks.test(x=residuals(morepower11), y=rnorm(200, sd=sqrt(0.3178622)))

        Two-sample Kolmogorov-Smirnov test

data:  residuals(morepower11) and rnorm(200, sd = sqrt(0.3178622))
D = 0.255, p-value = 4.498e-06
alternative hypothesis: two-sided

> set.seed(20230530)
> ks.test(x=residuals(morepower11), y=rnorm(200, sd=sqrt(0.3178622)))

        Two-sample Kolmogorov-Smirnov test

data:  residuals(morepower11) and rnorm(200, sd = sqrt(0.3178622))
D = 0.195, p-value = 0.0009959
alternative hypothesis: two-sided
```

Differences between the two tests

- S-W test is a modification of the Kolmogorov-Smirnov (K-S test) and gives more weight to the tails of the distribution than does the K-S test.
- The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested.

In class Practice Problem 18

From the clerical staff work hours, use residual plots to conduct a residual analysis of the data. **Check Normality Assumption by graphs, Shapiro-Wilk normality test and Kolmogorov-Smirnov test**. If you detect a trend, how would you like to transform the predictors in the model?

☞ Lecture_6.R

Take away messages

	Assumptions	Descriptions	Measurements	Potential actions
1	Linearity Assumption	Relationships between all predictors and the response are linear	1. residuals – fitted y values plot	Add high order terms or transform data
2	Independence Assumption	Independence of observations	1. Residuals vs predictors	Use other models than MLR
3	Equal Variance Assumption	Error term has equal variance given any values of independent variables	1. Scale-Location plot. [square root of standardized residuals – fitted y values plot] 2. Breusch-Pagan test (H0: no hetero)	Add high order terms or transform data
4	Normality Assumption	Error term is normally distributed	1. Histogram Normal probability plot 2. Q-Q plot 3. Kolmogorov-Smirnov test 4. Shapiro-Wilk test	Add high order terms or transform data
5	Multicollinearity			
6	Outlier			

Thank you

- Questions OR Comments?

- Slack channel: section2-course-documents
- Email: qing.li2@uclagary.ca