Linear Statistical Models with R

Modeling Teenage Gambling Expenditures

Andrew Craun   Mentor: Pei Zhang

University of Maryland, College Park

May 5, 2021

Outline

- Introduction to Linear Models
- Introduction to Least Squares Estimation
- Introduction to Our Data Set
- Checking Error Assumptions
- Variable Selection
- Summary and Concluding Remarks

Introduction to Linear Models: What is a Linear Model?

- A linear model is an equation that tells us the relationship between a response and some predictors (variables).
- We will try to find the coefficients β_1 ... β_n for each predictor variable x_1 ... x_n, as well as the intercept term β_0. This will help us solve for μ, which represents the expected value of the response variable y. A linear model's equation will look like this:
- $\mu = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ or $\mu = X\beta$
- When we are talking about the actual observed data, we add an error term ϵ to account for the fact that our observed data may be different from out predicted value μ:
- $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon_n$ or $\mathbf{y} = X\beta + \epsilon$

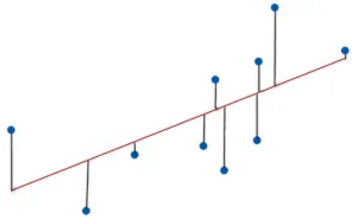
Introduction to Least Squares Estimation

- When making our linear model, we will use least squares estimation. This means that we want to minimize the sum of squares of the residuals (the observed values minus the fitted values).

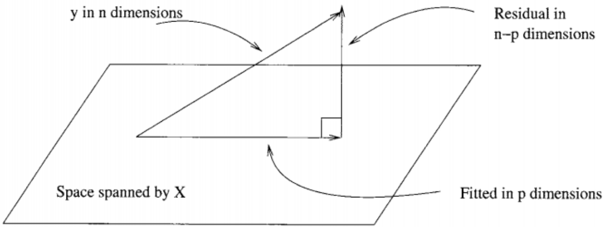
$$\sum \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

Introduction to Least Squares Estimation

- Residuals in 2 dimensions:



- Residuals in more dimensions:



Introduction to Our Data Set

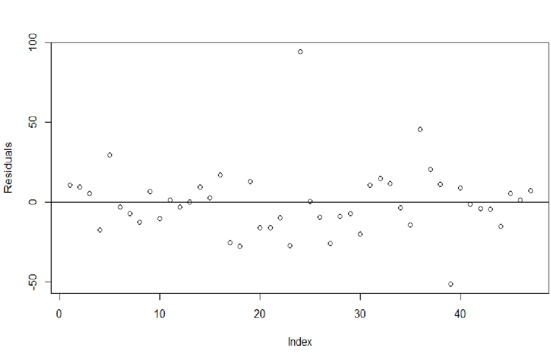
- For this presentation, I will demonstrate the linear regression methods using the teengamb dataset in R's faraway package. It was a survey that was conducted to study teenage gambling in Britain (Ide-Smith Lea, 1988, Journal of Gambling Behavior, 4, 110-118).
- The response variable is ”gamble”, which is expenditure on gambling in pounds per year.
- The predictors are:
- ”sex” = sex, 0 = ”male”, 1 = ”female”
- ”status” = socioeconomic status score based on parents' occupation
- ”income” = income in pounds per week
- ”verbal” = verbal score in words out of 12 correctly defined
- We will also add interaction terms between each of these predictors, to account for the fact that some of them may influence each other.

Checking Error Assumptions

- The first step to making a linear model is to make sure that a linear model is actually suitable for the data.
- To do this, we must first check the error assumptions: independence, constant variance, and normality of the errors.

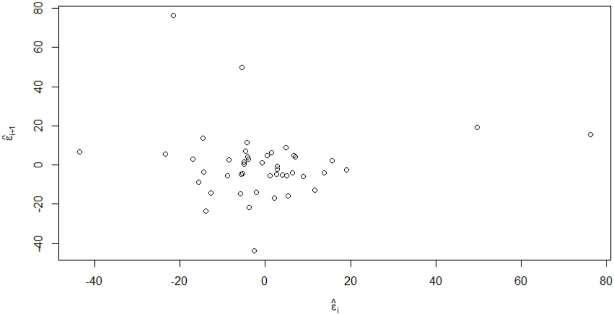
Independence of Errors

- Independence of errors means that the errors do not influence each other.
- To check, we look at the residuals plot. We look to see that the residuals don't have any pattern as we go from one person to the next.



Independence of Errors

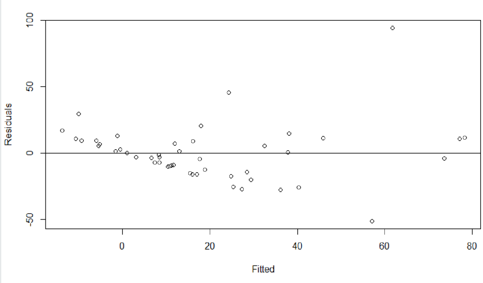
- We can also look at a successive residuals plot, which plots the residual on the horizontal axis and the residual of the next person in the survey on the vertical axis. If there is any pattern, such as the points going from the bottom left to top right, our assumption may be broken.



- These plots have no pattern, so we have independence of the errors.

Constant Variance of Errors

- Next, we check for constant variance of the errors. This means that for different areas of fitted values, the residuals are consistent. We can verify this by looking at a residuals vs. fitted plot.

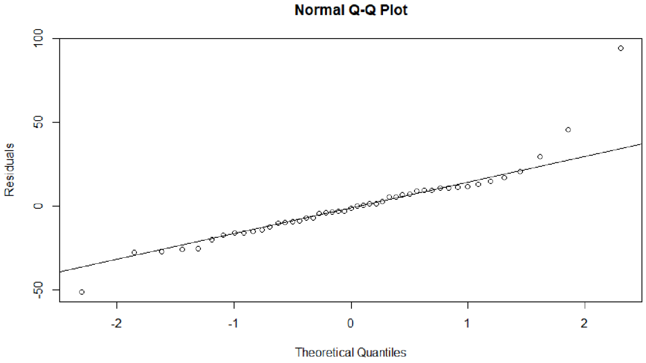


- We don't see any patterns of larger or smaller residuals along the fitted line, so we have constant variance of the errors.

Errors Normally Distributed

- Last, we must check for normal distribution of the errors. When taking a large sample, in our case 47 people, the error terms should end up being similar to a normal distribution due to the central limit theorem. To check for this, we look at a Q-Q plot.
- A Q-Q plot puts quantiles of a standard normal distribution on the horizontal axis and the residuals on the vertical axis.

Errors Normally Distributed



- Since the residuals approximately fit the normal distribution, we have normally distribution of the errors.
- Now that we have satisfied the three error assumptions, we can start to make our linear model.

Variable Selection

- When making a linear model, it is important to choose appropriate predictor variables to analyze.
- We ideally want to pick the smallest number of predictor variables to look at.
- Adding variables that aren't related to the response variable messes with the correlation between the more relevant predictors and the response.
- Also, measuring less predictors will make the experiment cost less money and take less time.

Variable Selection

- The simplest way to select variables is through backwards elimination. We start with the full model and look at the p value for each predictor, with the null hypothesis that the predictor has no effect on the response. Then, we remove the predictor with the highest p value and refit the linear model.
- Another method of variable selection is maximizing the adjusted R^2:

$$R_a^2 = \left(1 - \frac{RSS/(n-p)}{TSS/(n-1)}\right) \tag{1}$$

- where n = the sample size and p = the number of predictors. Unlike R^2, R_a^2 penalizes you for increasing the number of predictors, which helps you find a balance between a good fit and having too many predictors.

Variable Selection

- Similar to R_a^2, Mallow's C_p statistic is a criterion for balancing good fit with number of predictors:

$$C_p = \left(\frac{RSS_p}{\hat{\sigma}^2} + 2p - n \right) \tag{2}$$

- where $RSS_p =$ the the RSS of the model with p predictors and $\hat{\sigma}^2$ is from the model with all predictors. Unlike R_a^2, we want to minimize the C_p statistic rather than maximize it.

Variable Selection

- We can also use criterion-based procedures. One of these is the Akaike Information Criterion (AIC). The AIC value tells us how well a model fits the data. We want to find the number of parameters that minimizes the AIC value.

$$AIC = (2p - 2logL) \tag{3}$$

  where L is the likelihood function and p is the number of parameters.

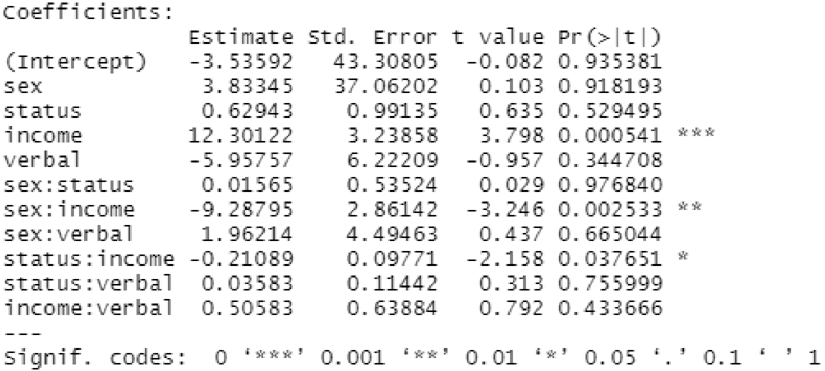
- Another criterion-based procedure is the Bayes Information Criterion (BIC):

$$BIC = (plogn - 2logL) \tag{4}$$

- For linear regression models, the $-2logL$ is nlog(RSS/n) + a constant.

Finding the Best Model for teengamb

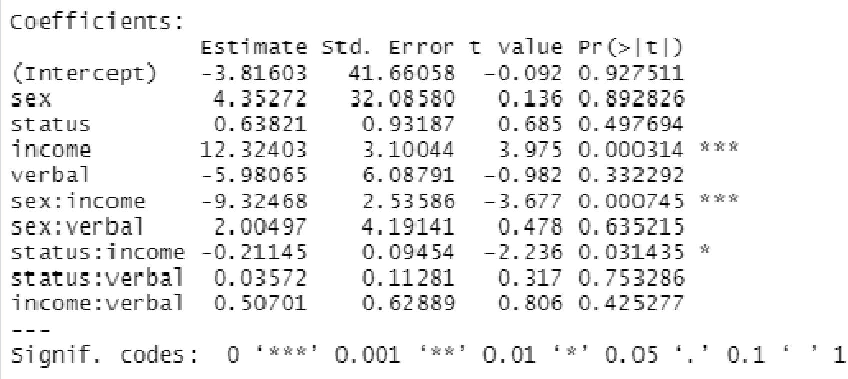
- Now, let's find the best model for our dataset. First, we will use backward elimination. The p-values for the variables are:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -3.53592   43.30805  -0.082 0.935381    
sex             3.83345   37.06202   0.103 0.918193    
status          0.62943    0.99135   0.635 0.529495    
income         12.30122    3.23858   3.798 0.000541 ***
verbal         -5.95757    6.22209  -0.957 0.344708    
sex:status      0.01565    0.53524   0.029 0.976840    
sex:income     -9.28795    2.86142  -3.246 0.002533 ** 
sex:verbal      1.96214    4.49463   0.437 0.665044    
status:income  -0.21089    0.09771  -2.158 0.037651 *  
status:verbal   0.03583    0.11442   0.313 0.755999    
income:verbal   0.50583    0.63884   0.792 0.433666    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We can see that the interaction term between sex and status has the highest p value, so we remove it and refit the model.

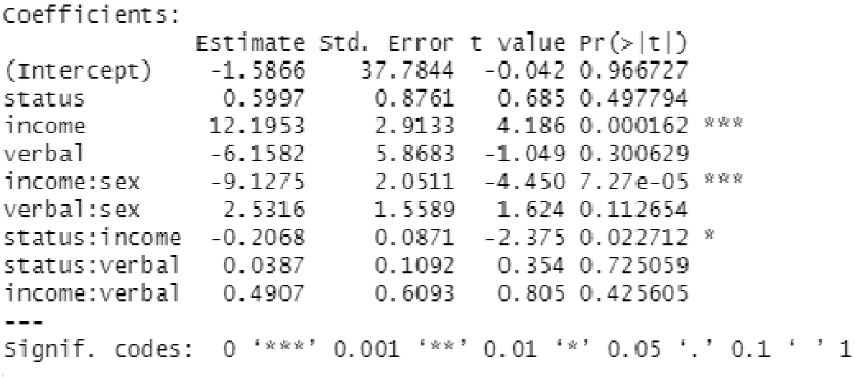
Finding the Best Model for teengamb



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -3.81603   41.66058  -0.092 0.927511    
sex             4.35272   32.08580   0.136 0.892826    
status          0.63821    0.93187   0.685 0.497694    
income         12.32403    3.10044   3.975 0.000314 ***
verbal         -5.98065    6.08791  -0.982 0.332292    
sex:income     -9.32468    2.53586  -3.677 0.000745 ***
sex:verbal      2.00497    4.19141   0.478 0.635215    
status:income  -0.21145    0.09454  -2.236 0.031435 *  
status:verbal   0.03572    0.11281   0.317 0.753286    
income:verbal   0.50701    0.62889   0.806 0.425277    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Now, sex has the highest p value, so we remove it.

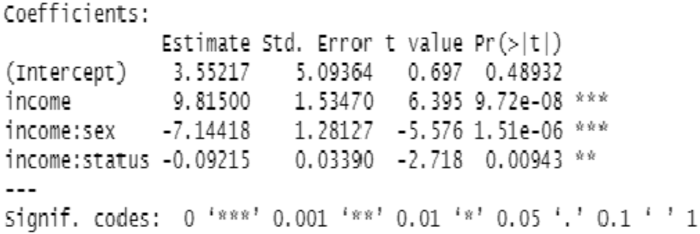
Finding the Best Model for teengamb



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -1.5866    37.7844  -0.042 0.966727    
status           0.5997     0.8761   0.685 0.497794    
income          12.1953     2.9133   4.186 0.000162 ***
verbal          -6.1582     5.8683  -1.049 0.300629    
income:sex      -9.1275     2.0511  -4.450 7.27e-05 ***
verbal:sex       2.5316     1.5589   1.624 0.112654    
status:income   -0.2068     0.0871  -2.375 0.022712 *  
status:verbal    0.0387     0.1092   0.354 0.725059    
income:verbal    0.4907     0.6093   0.805 0.425605    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We continue this process until we are left with predictors that are significant at the 0.05 level.

Finding the Best Model for teengamb

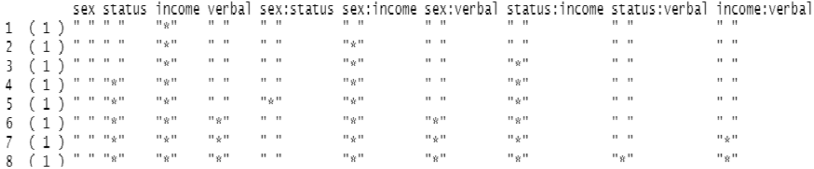


```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     3.55217    5.09364   0.697  0.48932    
income          9.81500    1.53470   6.395 9.72e-08 ***
income:sex     -7.14418    1.28127  -5.576 1.51e-06 ***
income:status  -0.09215    0.03390  -2.718  0.00943 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- So backwards elimination tells us that our final model should include income, income:sex, and income:status as predictors.

Finding the Best Model for teengamb

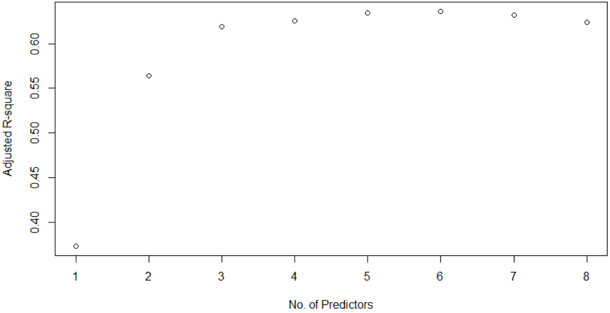
- Now, let's use R_a^2.
- To compare the models of different predictors, we must first find the best model for each number of predictors. R can do this by computing the RSS of each possible model for up to 8 predictors.



```
         sex status income verbal sex:status sex:income sex:verbal status:income status:verbal income:verbal
1  ( 1 ) " " " "    "*"    " "    " "        " "        " "        " "           " "           " "
2  ( 1 ) " " " "    "*"    " "    " "        "*"        " "        " "           " "           " "
3  ( 1 ) " " " "    "*"    " "    " "        "*"        " "        "*"           " "           " "
4  ( 1 ) " " "*"    "*"    " "    " "        "*"        " "        "*"           " "           " "
5  ( 1 ) " " "*"    "*"    " "    "*"        "*"        " "        "*"           " "           " "
6  ( 1 ) " " "*"    "*"    "*"    " "        "*"        "*"        "*"           " "           " "
7  ( 1 ) " " "*"    "*"    "*"    " "        "*"        "*"        "*"           " "           "*"
8  ( 1 ) " " "*"    "*"    "*"    " "        "*"        "*"        "*"           "*"           "*"
```

- This R output tells us the best predictors for a model with a set number of predictors.
- Now, we calculate the R_a^2 for each model.

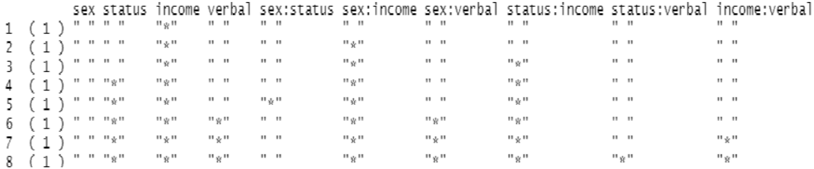
Finding the Best Model for teengamb



- We can see that the R_a^2 is highest at 6 predictors, so we definitely don't want more than 6. Now, we must make a decision about the trade off between increasing R_a^2 slightly and adding another predictor. Between 3 and 6 predictors, the R_a^2 increase is very small, meaning adding one more predictor is not as significant. So based on R_a^2 , we should consider the 3, 4, 5 , and 6 predictor models.

Finding the Best Model for teengamb

- Looking at the chart of best models from earlier:

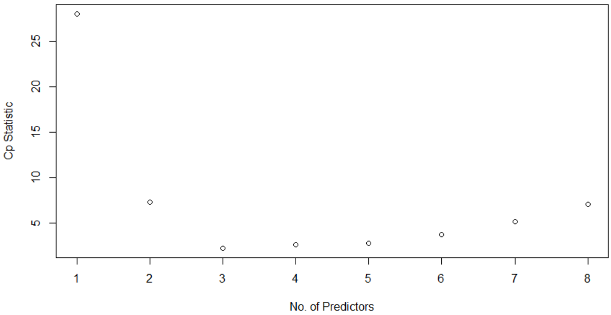


```
         sex status income verbal sex:status sex:income sex:verbal status:income status:verbal income:verbal
1  ( 1 ) " " " "    "*"    " "    " "        " "        " "        " "           " "           " "
2  ( 1 ) " " " "    "*"    " "    " "        "*"        " "        " "           " "           " "
3  ( 1 ) " " " "    "*"    " "    " "        "*"        " "        "*"           " "           " "
4  ( 1 ) " " "*"    "*"    " "    " "        "*"        " "        "*"           " "           " "
5  ( 1 ) " " "*"    "*"    " "    "*"        "*"        " "        "*"           " "           " "
6  ( 1 ) " " "*"    "*"    "*"    " "        "*"        "*"        "*"           " "           " "
7  ( 1 ) " " "*"    "*"    "*"    " "        "*"        "*"        "*"           " "           "*"
8  ( 1 ) " " "*"    "*"    "*"    " "        "*"        "*"        "*"           "*"           "*"
```

- The 3 parameter model should include income, sex:income, and status:income.
- This agrees with our result from backward elimination, which we love to see.
- If we were to use the four, five or six predictor models, we would use the corresponding predictors.

Finding the Best Model for teengamb

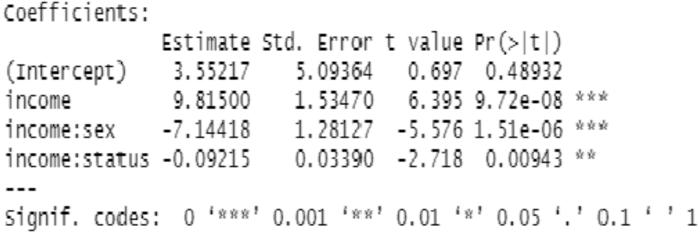
- We can also take a look at Mallow's C_p statistic to help with our conclusion:



- Here, the three parameter model is best, so we should go with the three predictor model.

Summary

- So, after verifying that our data fits the requirements of a linear model and then narrowing down the predictors, our final model is



```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.55217    5.09364   0.697  0.48932
income           9.81500    1.53470   6.395 9.72e-08 ***
income:sex      -7.14418    1.28127  -5.576 1.51e-06 ***
income:status   -0.09215    0.03390  -2.718  0.00943 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

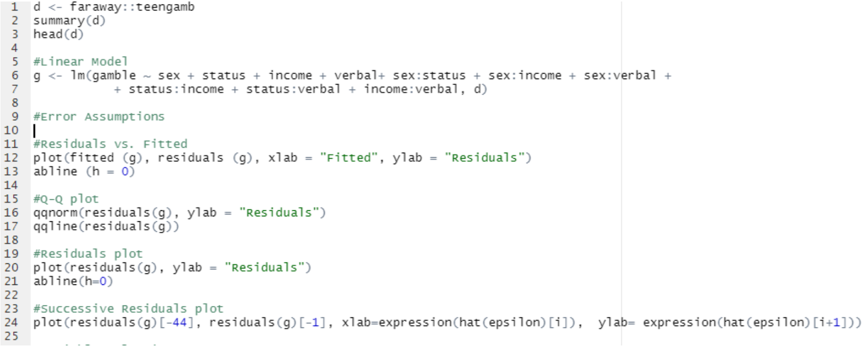
- gamble $=$ 0.48932 $+$ 9.81500(income) - 7.14418(income:sex) - 0.09215(income:status)
- What does this tell us?

Reference

- Agresti, A. (2015). Foundations of Linear and Generalized Linear Models. Hoboken, NJ: John Wiley & Sons.
- Faraway, J. J. (2004). Linear Models with R. Boca Raton, FL: Chapman & Hall/CRC.
- Images:
- https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

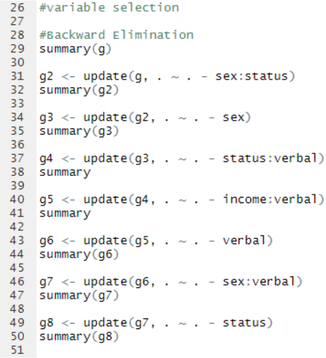
Thanks

Appendix



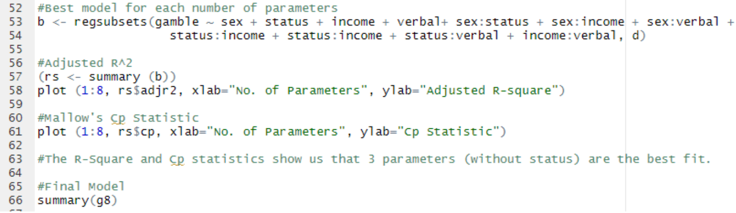
```r
d <- faraway::teengamb
summary(d)
head(d)

#Linear Model
g <- lm(gamble ~ sex + status + income + verbal+ sex:status + sex:income + sex:verbal +
           + status:income + status:verbal + income:verbal, d)

#Error Assumptions

#Residuals vs. Fitted
plot(fitted (g), residuals (g), xlab = "Fitted", ylab = "Residuals")
abline (h = 0)

#Q-Q plot
qqnorm(residuals(g), ylab = "Residuals")
qqline(residuals(g))

#Residuals plot
plot(residuals(g), ylab = "Residuals")
abline(h=0)

#Successive Residuals plot
plot(residuals(g)[-44], residuals(g)[-1], xlab=expression(hat(epsilon)[i]),  ylab= expression(hat(epsilon)[i+1]))

```

Appendix



```
#variable selection

#Backward Elimination
summary(g)

g2 <- update(g, . ~ . - sex:status)
summary(g2)

g3 <- update(g2, . ~ . - sex)
summary(g3)

g4 <- update(g3, . ~ . - status:verbal)
summary

g5 <- update(g4, . ~ . - income:verbal)
summary

g6 <- update(g5, . ~ . - verbal)
summary(g6)

g7 <- update(g6, . ~ . - sex:verbal)
summary(g7)

g8 <- update(g7, . ~ . - status)
summary(g8)

```

Appendix



```
#Best model for each number of parameters
b <- regsubsets(gamble ~ sex + status + income + verbal+ sex:status + sex:income + sex:verbal +
                  status:income + status:income + status:verbal + income:verbal, d)

#Adjusted R^2
(rs <- summary (b))
plot (1:8, rs$adjr2, xlab="No. of Parameters", ylab="Adjusted R-square")

#Mallow's Cp Statistic
plot (1:8, rs$cp, xlab="No. of Parameters", ylab="Cp Statistic")

#The R-Square and Cp statistics show us that 3 parameters (without status) are the best fit.

#Final Model
summary(g8)
```