STP 429 – Applied Regression

Arizona State University

Tanay Gandhi

Lab #3

Executive Summary

Career stock market investors and retail investors alike rely heavily on the ability to predictive short term and long term growth and losses of companies on the market to make profit. Being able to predict future stock gains means big potential monetary wins. An industry of interest, particularly in the last year after the initial market crash from COVID-19 lockdowns in 2020 is the travel industry. Company stocks in the travel industry took a big hit, as most people were completely restricted from travelling. This, however, also spelled potential wins for investors, as many travel companies have a resilient balance sheet and could survive through the pandemic.

Marriott International Inc. is a company on the S&P 500 index that owns and operates many large hotel and residential brands all over the world. Well known hotel brands under Marriott include Courtyard, Ritz-Carlton, Westin, and AC hotels. They offer business to a very wide variety of customers due to their diverse investments and ownerships. They own hotel brands geared towards business travel as well as leisure. Because of this, the company has a wide variety of incomes. Furthermore, the company's balance sheets in recent years have show larger and more stable growth, which allows many investors to view this company as a stable but lucrative investment. This study aims to develop a predictive model using other company stock prices on the S&P 500 index.

Methodology

The initial independent variables were chosen by analyzing the correlation between each independent variable in the dataset and MAR. The decision to begin the analysis with each variable was made based on the correlation between that variable and MAR, as well as existing knowledge of the relationship between the two companies. Once the initial five variables were chosen, their histograms and distribution metrics were analyzed to identify possible outliers, skew, and potential for interaction. Next, correlation analysis was performed between each of the five variables and Marriott to further understand the relationships between these variables. Scatterplots were used to identify potential benefits of quadratic term usage in the final model. Correlation analysis between independent variables was used to identify potential for multicollinearity. Then, stepwise regression was performed to identify the most relevant parameters for the final model. Then, the variance inflation factors of each variable in the model were analyzed to possible multicollinearity in the model.

All processes were accomplished using SAS statistical software. Exploratory data analysis using the `proc means`, `proc corr`, and `proc univariate` procedures. Correlation analysis was completed using the `proc corr` procedure. Finally, the various regression models were generated using the `proc reg` and `proc glmselect` procedures.

Results

In order to identify the initial independent variables to begin this study, I analyzed the correlation coefficients between each of the 61 independent variables and Marriott, to identify variables having strong relationships with the Marriott data. An understanding of the relationship between hotel chains and companies in adjacent commercial sectors, such as travel planners, trip aggregators, and cruise lines was also used to identify five variables with the strongest potential to benefit a predictive model for the price of Marriott stock. Table 1 shows us that the BKNG, EXPE, HLT, MGM, and NCLH symbols had the highest correlation with Marriott price, and this aligns with my initial prediction. Booking Holdings (BKNG) and Expedia (EXPE) are technology companies that provide travel aggregator services for consumer travel. Hilton (HLT) and MGM Resorts International (MGM) are both hospitality companies that own both destination resorts and multiple hotel chains. Finally, Norwegian Cruise Line Holdings is an American Cruise Line. All of these identified companies participate in the more general travel industry, so it is expected that they have a high correlation with Marriott stock price and were chosen to be analyzed further in this study.

Figure 1 shows that the distribution of Marriott stock price has two distinct modes, one centered around $96.00 and another around $132.00. Considering the stock market throughout the COVID-19 pandemic, it is expected that many of these companies will have a somewhat bimodal distribution. The mode centered around the smaller value could be the price after the initial market crash in March of 2020, as the American and world economies remained at a standstill for some time during the initial lockdown. The second mode is very apparent with Marriott but may not be so apparent in other distributions and may be the result of the current state of the travel industry, in which travel is somewhat limited and travel companies have little room for growth. MAR did not have any outliers (table 3). The Booking Holdings distribution, as seen in figure 2, has a surprisingly unimodal distribution, centered around $1725.00, with a slight right skew. Interestingly, the middle 80% of the data evenly occurred at around the same frequency, about 10-15%, which possibly indicates low variability in price and more gradual growth/loss in stock price. BKNG did not have any outliers (table 4). Expedia’s distribution (figure 3) is bimodal, with a mode around $82.50 and another around $127.50. The lack of shape possibly indicates high fluctuation in EXPE price, which is interesting compared to BKNG, as they both participate in the travel aggregator sector. EXPE does not have any outliers (table 5). Hilton’s distribution in figure 4 can be described as bimodal with a mode around $88.00 and another around $104.00. Prices between $72.00 and $88.00 interestingly occurred in over 50% of the dataset, which is significant and is possibly due to low fluctuations in that price range. HLT did not have any outliers (table 6). The distribution of MGM Resorts International in figure 5 had very little shape and a mode around $22.50, with large grouping occurring around $16.50, $31.50, and $37.50. The lack of shape in this distribution indicates large fluctuations in price, and this makes sense considering the timeline. The previously mentioned companies were able to maintain some stable business due to the relatively quick resurgence of business travel. MGM, however, specializes in destination hotels and rely almost entirely on travel for pleasure, which continues to be limited during the pandemic. MGM did not have any outliers (table 7). Finally, the Norwegian stock price distribution in figure 6 is somewhat bimodal, with a mode at around $15.00 and another large grouping around $24.00. NCLH does not have any outliers (table 8).

Revisiting table 1, all of the identified variables have very strong correlation with MAR, each having a correlation coefficient >0.96. HLT has the highest correlation with MAR, which is expected, as both exist as equal competitors in the hospitality industry and both own hotel chains that service business travelers and vacationers equally. MGM has the next highest correlation, which is interesting considering the high fluctuation of its distribution but is expected as it also contributes to and is affected by the hospitality industry. The scatter plot in figure 10 shows a possibility for quadratic term usage. Interestingly, the significant groupings noticed in the distributions of EXPE and MGM reappear in their scatter plots against MAR in both dimensions, indicating that Marriott's price experienced fluctuation and stagnation at a similar rate to EXPE and MGM. NCLH had the lowest correlation with MAR and the scatter plot in figure 11 reflects this, as it is much more varied than the other plots. It does, however, contain two significant groupings, which does show that NCLH price had some significant relation to MAR. The plot's shape also indicates a possible need for quadratic terms.

Finally, after identifying and analyzing the independent variables BKNG, EXPE, HLT, MGM, and NCLH, stepwise regression was performed to select the most useful variables for a base first-order model. The process took three steps and selected the variables HLT, NCLH, and EXPE. HLT was the first predictor picked by the stepwise regression procedure, which makes sense because it is another hotel company. Interestingly, MGM was not one of the variables chosen, even though it is another hotel-related company and had a higher correlation with MAR than EXPE, which was chosen. This may be due to the fact that it might not offer enough new information and benefit the model significantly since HLT is already included. In the R^2 selection method, the model $MAR = \beta_0 + \beta_1 * EXPE + \beta_2 * HLT + \beta_3 * NCLH$ contained the highest R^2 value of all possible 3-variables models at 0.9798, which is only 0.0002 behind the model that best covers data variation, the one that contained all five predictors. The adjusted R^2 selection procedure had similar results. The adjusted R^2 of the $MAR = \beta_0 + \beta_1 * EXPE + \beta_2 * HLT + \beta_3 * NCLH$ model was 0.9795. Once again, this was the highest out of all models with 3 variables and under, and only beaten by a few 4 variable models and the full five variable model. The results of the C_p criterion support the original stepwise regression. The model with the smallest C_p was $MAR = \beta_0 + \beta_1 * EXPE + \beta_2 * HLT + \beta_3 * NCLH$. Unlike the r^2 and r^2 adjusted criterion procedures, this model beat out the four variable and five variable models. This makes sense, as the C_p of a model is greatly increased with each additional parameter. Because of this, the model with the least number of variables that also has the least amount of regression bias is the one that includes EXPE, HLT, and NCLH. Finally, the best model determined by the PRESS criterion method was once again $MAR = \beta_0 + \beta_1 * EXPE + \beta_2 * HLT + \beta_3 * NCLH$. Since regular stepwise regression, C_p, and PRESS all picked this model, and it accounted for a very high percentage of variation in the data as shown in the r^2 and r^2 adjusted methods, this is a very good base predictive model. The final regressed base model (table 9) was:

$$MAR = 6.34509 + 0.08255 * EXPE + 0.741 * HTL + 1.22107 * NCLH$$

To confirm the usefulness of this base model, we will perform an f-test of utility:

$$H_0: All\ the\ coefficients = 0$$

$$H_\alpha: at\ least\ one\ coefficient\ \neq 0$$

The p-value result of this test is <0.0001 and with alpha=0.05, we reject the null hypothesis that all coefficients of the model are equal to 0.

The distributions of EXPE and HLT shown in figures 3 and 4 are very similarly shaped and this is also reflected in their respective scatter plots against MAR in figures 8 and 9. These scatter plots have remarkably similar shapes, with groupings of data points in similar areas. This leads to the possibility of some significant interaction between EXPE and HLT in the base model defined earlier. A new model with an interaction term between EXPE and HLT was regressed and the result was:

$$MAR = 23.667 - 0.2178 * EXPE + 0.64425 * HLT + 1.24442 * NCLH + 0.00222 * EXPE * HLT$$

It is immediately concerning that the sign of the coefficient of EXPE flipped from positive to negative. To test whether this new model is more useful than the original base model, we will use a nested f-test:

$$H_0: \beta_4 = 0$$

After confirming that interaction terms and quadratic terms are not statistically useful and adjusting for multicollinearity, the final predictive model for the price of MAR stock in this study is:

$$MAR = 33.70141 + 0.02251 * EXPE + 1.62172 * NCLH$$

The residual plot for this model shows a sufficient amount of scatter, although there are concerning groupings around the \$100 and \$130 predicted value mark, which may require further investigation.

Final Conclusion and Next Steps

This study was only performed with 61 initial independent variables from the S&P 500 index. This generated a fairly accurate and useful model, but there may be companies outside of this general industry that provide new and useful information for a more robust model. Specifically in the year 2020, the stock price of companies such as Johnson & Johnson and Pfizer may provide useful information, as the development and production of their vaccines directly benefitted the travel industry.

Appendix

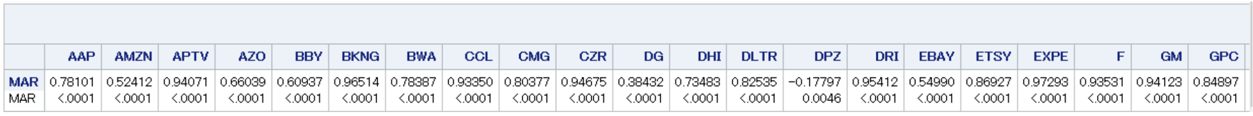
SAS Code

```
proc contents data=market;
run;

/* Picking initial variables*/

proc corr data=market;
var AAP AMZN APTV AZO BBY BKNG BWA CCL CMG CZR DG DHI DLTR DPZ DRI EBAY ETSY
EXPE F GM GPC GPS GRMN HAS HBI HD HLT KMX LB LEG LEN LKQ LOW LVS MCD MGM MHK
NCLH NKE NVR NWL ORLY PENN PHM POOL PVH RCL RL ROST SBUX TGT TJX TPR TSCO
TSLA UA ULTA VFC WHR WYNN YUM;
with MAR;
run;

/* Exploratory Data Analysis */
proc means data=market;
var MAR BKNG EXPE HLT MGM NCLH;
run;

proc univariate data=market;
var MAR BKNG EXPE HLT MGM NCLH;
histogram MAR BKNG EXPE HLT MGM NCLH;
run;

/* Correlation Analysis */
proc corr data=market plots=scatter();
var BKNG EXPE HLT MGM NCLH;
with MAR;
run;

/* Multiple Regression */
proc reg data=market;
model MAR = BKNG EXPE HLT MGM NCLH / selection=stepwise;
run;

/*R^2*/
proc reg data=market;
model MAR = BKNG EXPE HLT MGM NCLH / selection=RSQUARE;
run;

/*Adjusted R^2*/
proc reg data=market;
model MAR = BKNG EXPE HLT MGM NCLH / selection=ADJRSQ;
run;

/*CP Selection*/
proc reg data=market;
model MAR = BKNG EXPE HLT MGM NCLH / selection=CP;
run;

/*PRESS*/
```

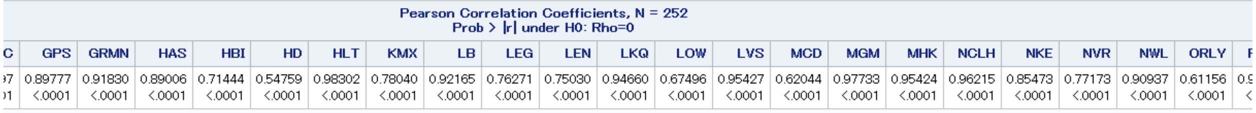
```
proc glmselect data=market;
model  MAR = BKNG EXPE HLT MGM NCLH /selection=stepwise(choose=press);
run;

/*Interaction Quad*/
data test_market;
set market;
interact = HLT * EXPE;
NCLH_sq = NCLH * NCLH;
run;

/*With interact term*/
proc reg data=test_market;
model MAR = EXPE HLT NCLH interact;
run;

/*With quad term*/
proc reg data=test_market;
model MAR = EXPE HLT NCLH NCLH_sq;
run;

/*Complete model*/
proc reg data=test_market;
model MAR = EXPE HLT NCLH NCLH_sq interact;
run;


/*Multicollinearity play*/
proc reg data=market;
model MAR = EXPE HLT NCLH / vif;
run;

/*Ridge plot*/
proc reg data=market outvif plots(only)=ridge(unpack VIFaxis=log);
model MAR = EXPE HLT NCLH / vif;
plot / ridgeplot nomodel nostat;
run;

/*Multicollinearity*/
proc corr data=market plots=scatter();
var EXPE HLT NCLH;
with EXPE HLT NCLH;
run;

proc reg data=market;
model MAR = EXPE NCLH / selection=stepwise vif;
run;

/*Complex models*/
data complete_mkt;
set market;
interact = HLT * EXPE;
NCLH_sq = NCLH * NCLH;
run;

proc reg data=market;
model MAR = EXPE HLT NCLH;
```

```
run;

/*Interaction*/
proc reg data=complete_mkt;
model MAR = EXPE HLT NCLH interact;
run;

/*Quadratic*/
proc reg data=complete_mkt;
model MAR = EXPE HLT NCLH NCLH_sq;
run;
```

SAS Output:

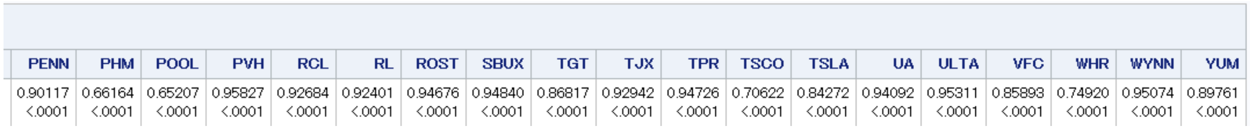


	AAP	AMZN	APTV	AZO	BBY	BKNG	BWA	CCL	CMG	CZR	DG	DHI	DLTR	DPZ	DRI	EBAY	ETSY	EXPE	F	GM	GPC
MAR MAR	0.78101 <.0001	0.52412 <.0001	0.94071 <.0001	0.66039 <.0001	0.60937 <.0001	0.96514 <.0001	0.78387 <.0001	0.93350 <.0001	0.80377 <.0001	0.94675 <.0001	0.38432 <.0001	0.73483 <.0001	0.82535 <.0001	−0.17797 0.0046	0.95412 <.0001	0.54990 <.0001	0.86927 <.0001	0.97293 <.0001	0.93531 <.0001	0.94123 <.0001	0.84897 <.0001



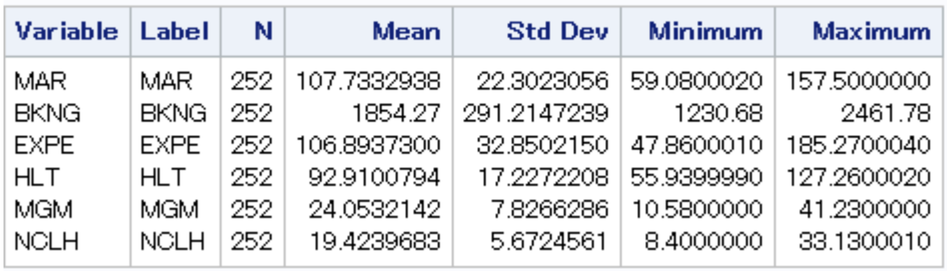
Pearson Correlation Coefficients, N = 252
Prob > |r| under H0: Rho=0

GPS	GRMN	HAS	HBI	HD	HLT	KMX	LB	LEG	LEN	LKQ	LOW	LVS	MCD	MGM	MHK	NCLH	NKE	NVR	NWL	ORLY
0.89777 <.0001	0.91830 <.0001	0.89006 <.0001	0.71444 <.0001	0.54759 <.0001	0.98302 <.0001	0.78040 <.0001	0.92165 <.0001	0.76271 <.0001	0.75030 <.0001	0.94660 <.0001	0.67496 <.0001	0.95427 <.0001	0.62044 <.0001	0.97733 <.0001	0.95424 <.0001	0.96215 <.0001	0.85473 <.0001	0.77173 <.0001	0.90937 <.0001	0.61156 <.0001



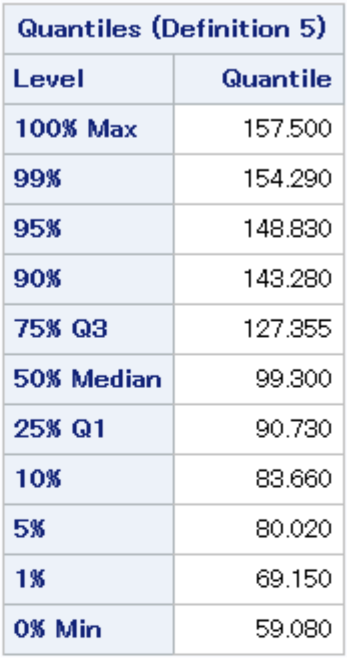
PENN	PHM	POOL	PVH	RCL	RL	ROST	SBUX	TGT	TJX	TPR	TSCO	TSLA	UA	ULTA	VFC	WHR	WYNN	YUM
0.90117 <.0001	0.66164 <.0001	0.65207 <.0001	0.95827 <.0001	0.92684 <.0001	0.92401 <.0001	0.94676 <.0001	0.94840 <.0001	0.86817 <.0001	0.92942 <.0001	0.94726 <.0001	0.70622 <.0001	0.84272 <.0001	0.94092 <.0001	0.95311 <.0001	0.85893 <.0001	0.74920 <.0001	0.95074 <.0001	0.89761 <.0001

Table 1: Correlation matrix MAR vs all other variables



Variable	Label	N	Mean	Std Dev	Minimum	Maximum
MAR	MAR	252	107.7332938	22.3023056	59.0800020	157.5000000
BKNG	BKNG	252	1854.27	291.2147239	1230.68	2461.78
EXPE	EXPE	252	106.8937300	32.8502150	47.8600010	185.2700040
HLT	HLT	252	92.9100794	17.2272208	55.9399990	127.2600020
MGM	MGM	252	24.0532142	7.8266286	10.5800000	41.2300000
NCLH	NCLH	252	19.4239683	5.6724561	8.4000000	33.1300010

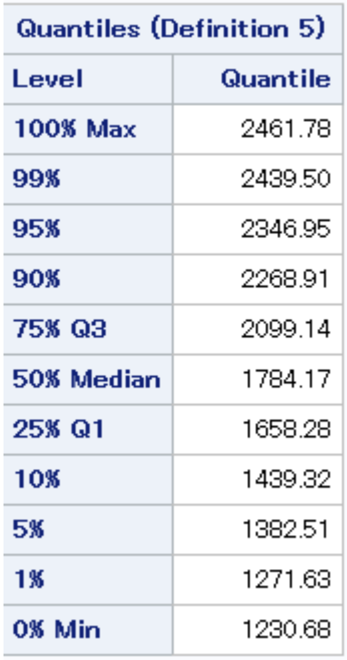
Table 2: Distribution metrics



Quantiles (Definition 5)

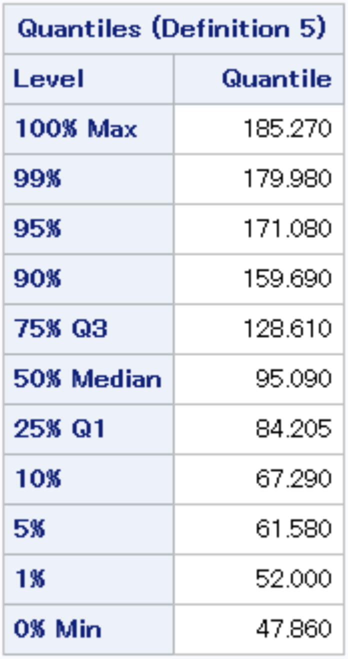
Level	Quantile
100% Max	157.500
99%	154.290
95%	148.830
90%	143.280
75% Q3	127.355
50% Median	99.300
25% Q1	90.730
10%	83.660
5%	80.020
1%	69.150
0% Min	59.080

Table 3: Quartiles of MAR



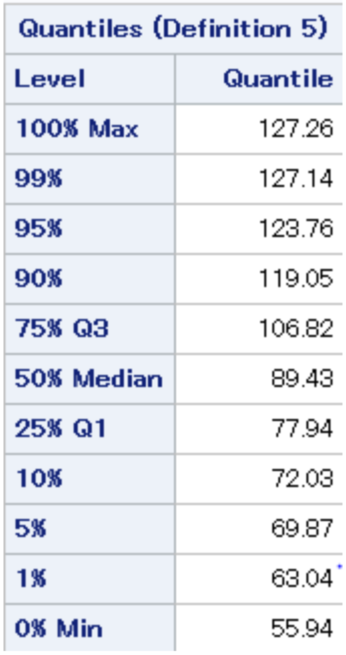
Quantiles (Definition 5)	
Level	Quantile
100% Max	2461.78
99%	2439.50
95%	2346.95
90%	2268.91
75% Q3	2099.14
50% Median	1784.17
25% Q1	1658.28
10%	1439.32
5%	1382.51
1%	1271.63
0% Min	1230.68

Table: 4 Quartiles of BKNG



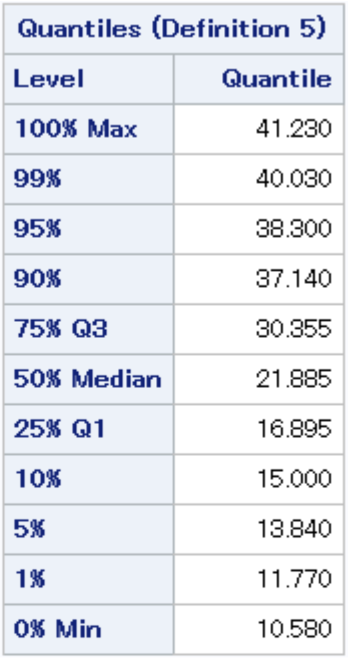
Quantiles (Definition 5)	
Level	Quantile
100% Max	185.270
99%	179.980
95%	171.080
90%	159.690
75% Q3	128.610
50% Median	95.090
25% Q1	84.205
10%	67.290
5%	61.580
1%	52.000
0% Min	47.860

Table 5: Quartiles of EXPE



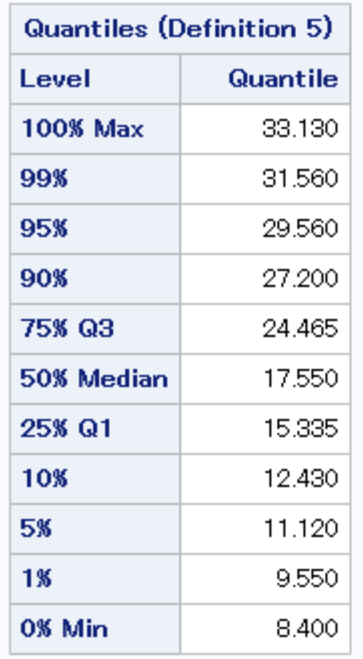
Quantiles (Definition 5)	
Level	Quantile
100% Max	127.26
99%	127.14
95%	123.76
90%	119.05
75% Q3	106.82
50% Median	89.43
25% Q1	77.94
10%	72.03
5%	69.87
1%	63.04
0% Min	55.94

Table 6: Quartiles of HLT



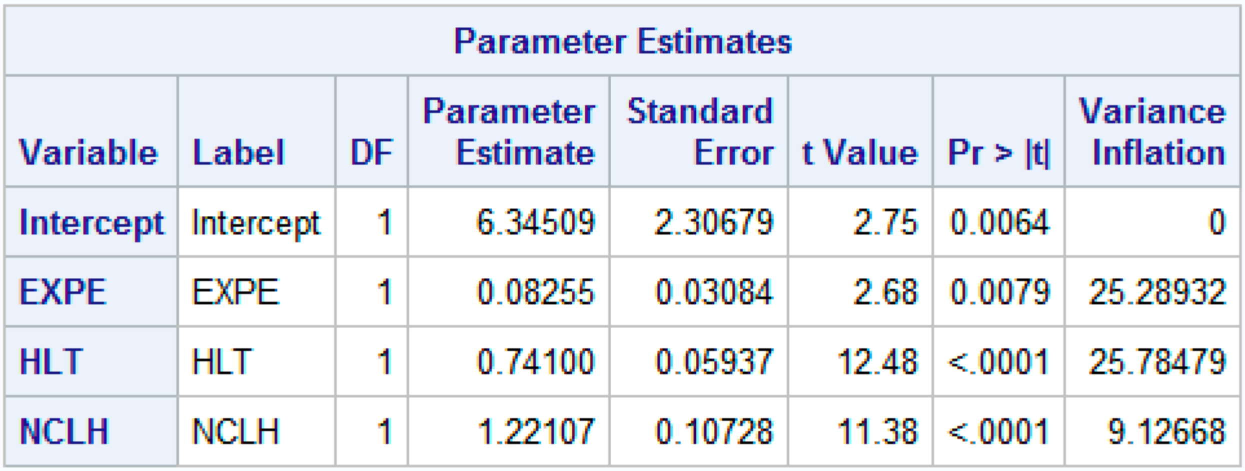
Quantiles (Definition 5)	
Level	Quantile
100% Max	41.230
99%	40.030
95%	38.300
90%	37.140
75% Q3	30.355
50% Median	21.885
25% Q1	16.895
10%	15.000
5%	13.840
1%	11.770
0% Min	10.580

Table 7: Quartiles of MGM



Quantiles (Definition 5)	
Level	Quantile
100% Max	33.130
99%	31.560
95%	29.560
90%	27.200
75% Q3	24.465
50% Median	17.550
25% Q1	15.335
10%	12.430
5%	11.120
1%	9.550
0% Min	8.400

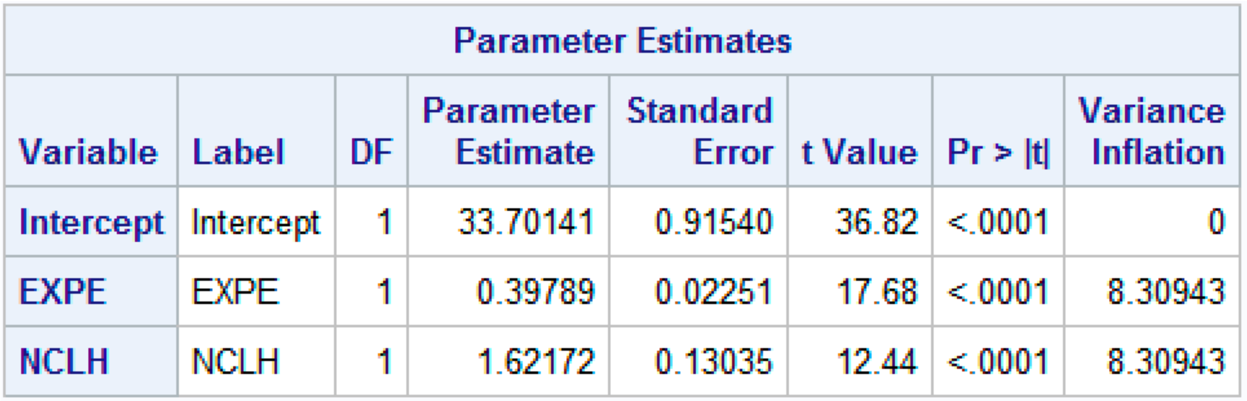
Table 8: Quartiles of NCLH



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	6.34509	2.30679	2.75	0.0064	0
EXPE	EXPE	1	0.08255	0.03084	2.68	0.0079	25.28932
HLT	HLT	1	0.74100	0.05937	12.48	<.0001	25.78479
NCLH	NCLH	1	1.22107	0.10728	11.38	<.0001	9.12668

Table 9: Regression of base model



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	33.70141	0.91540	36.82	<.0001	0
EXPE	EXPE	1	0.39789	0.02251	17.68	<.0001	8.30943
NCLH	NCLH	1	1.62172	0.13035	12.44	<.0001	8.30943

Table 10: Lower multicollinearity model

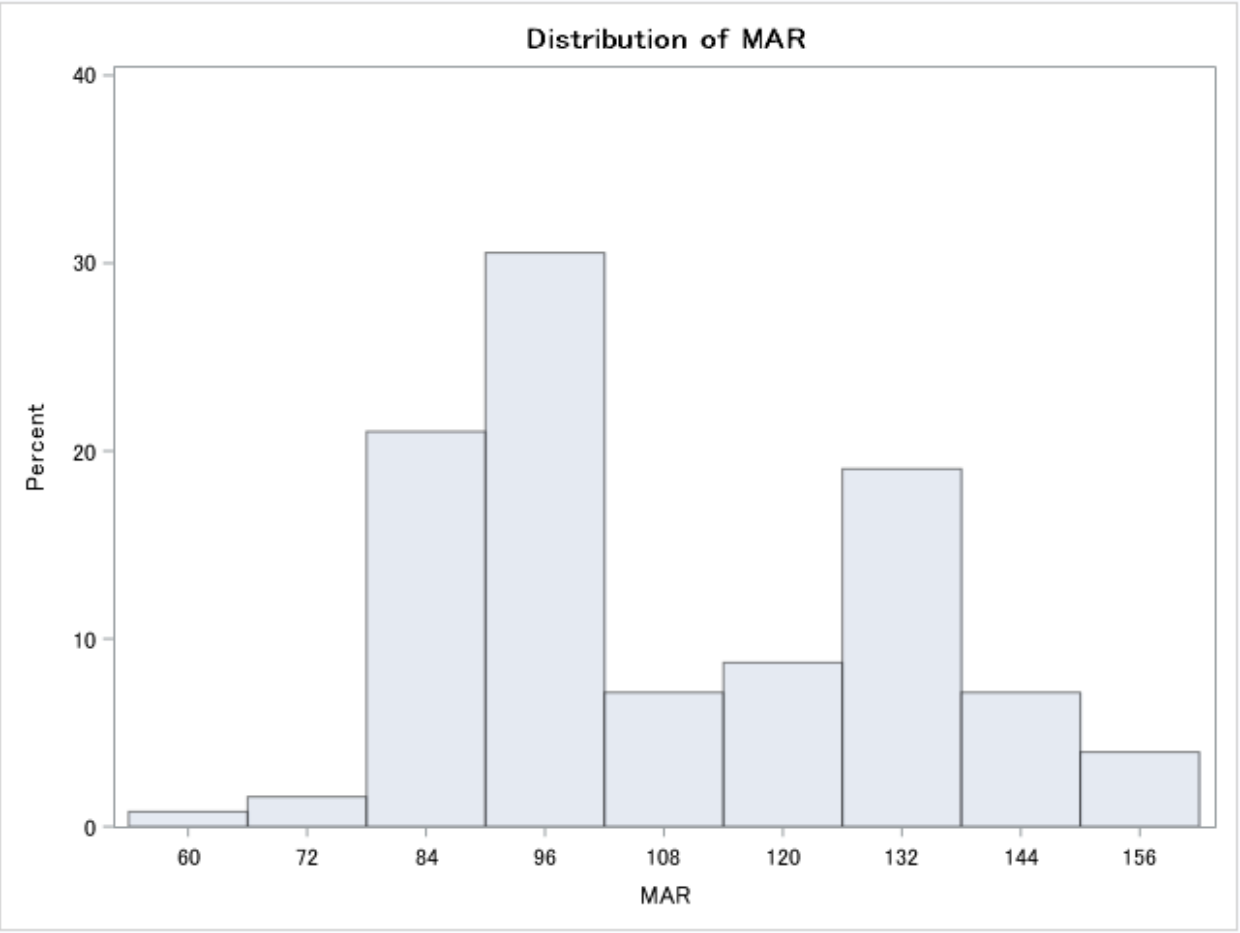


Figure 1: Distribution of MAR

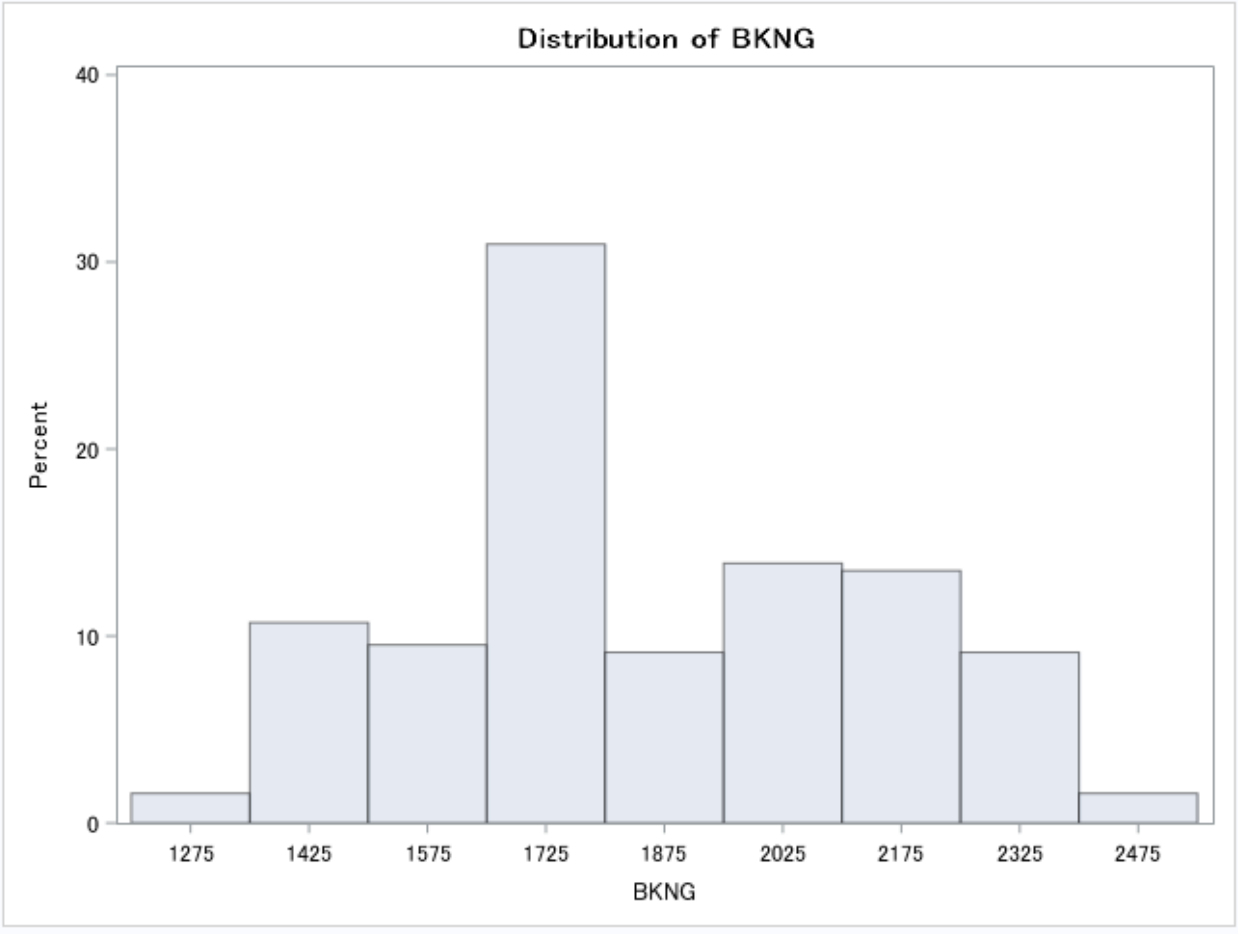


Figure 2: Distribution of BKNG

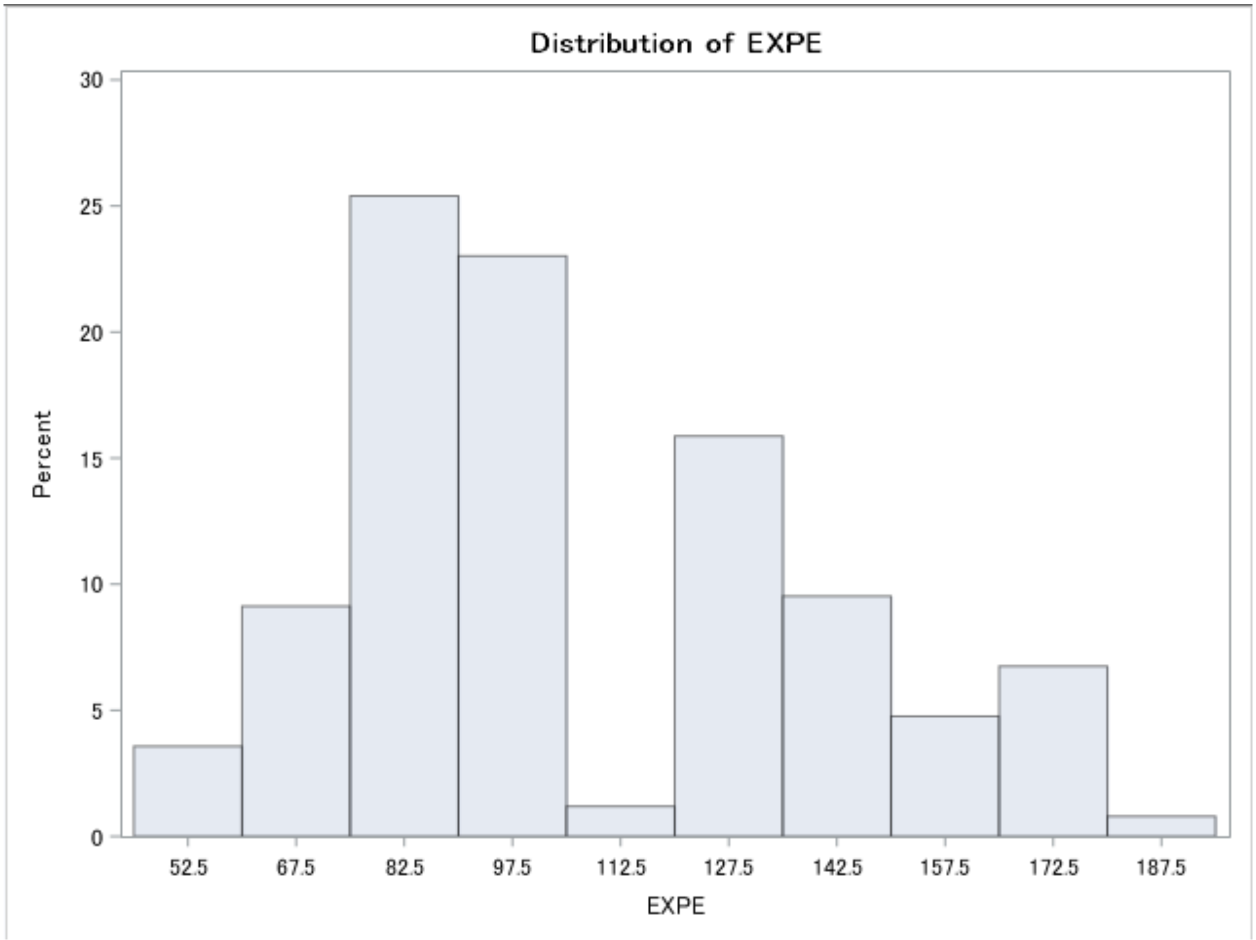


Figure 3: Distribution of EXPE

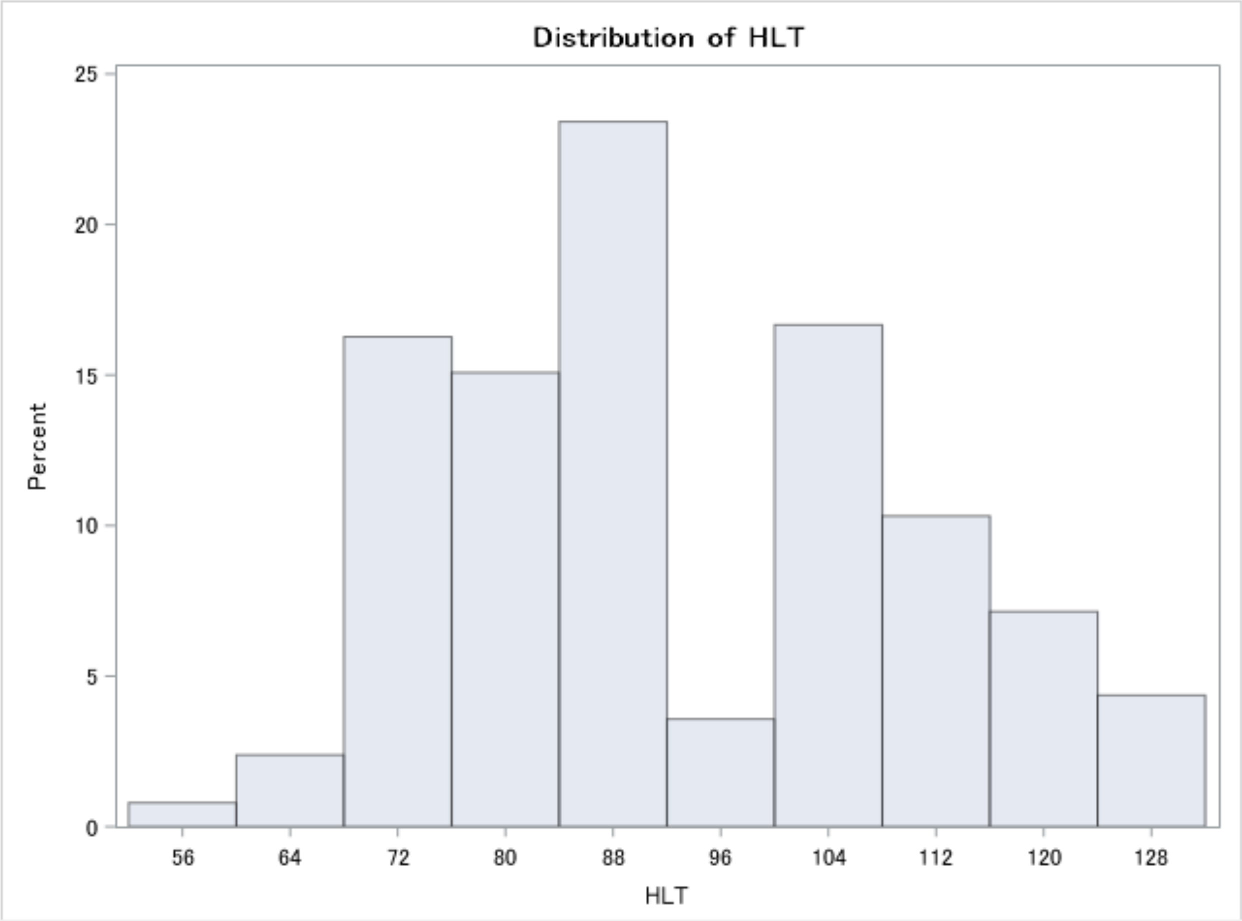


Figure 4: Distribution of HLT

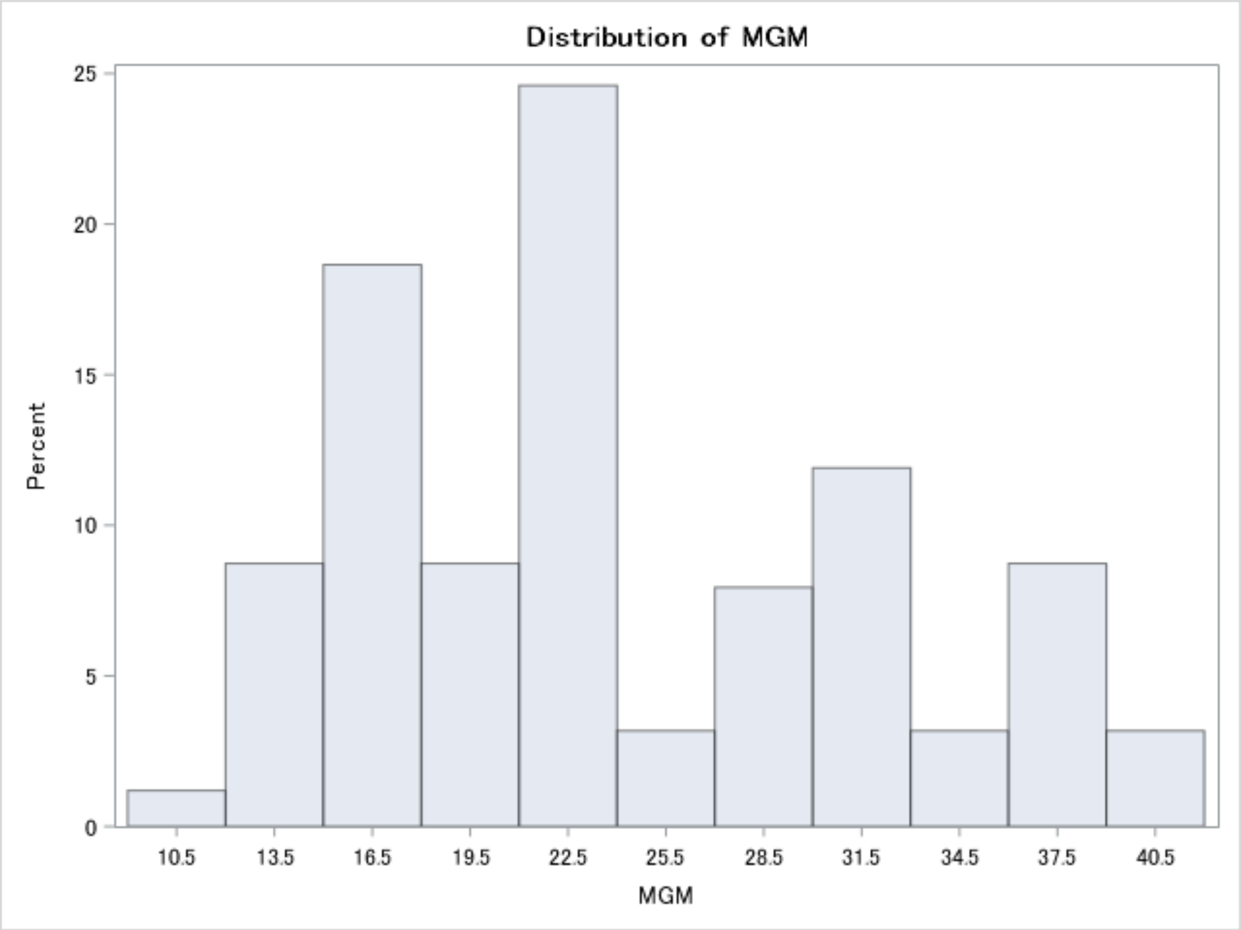


Figure 5: Distribution of MGM

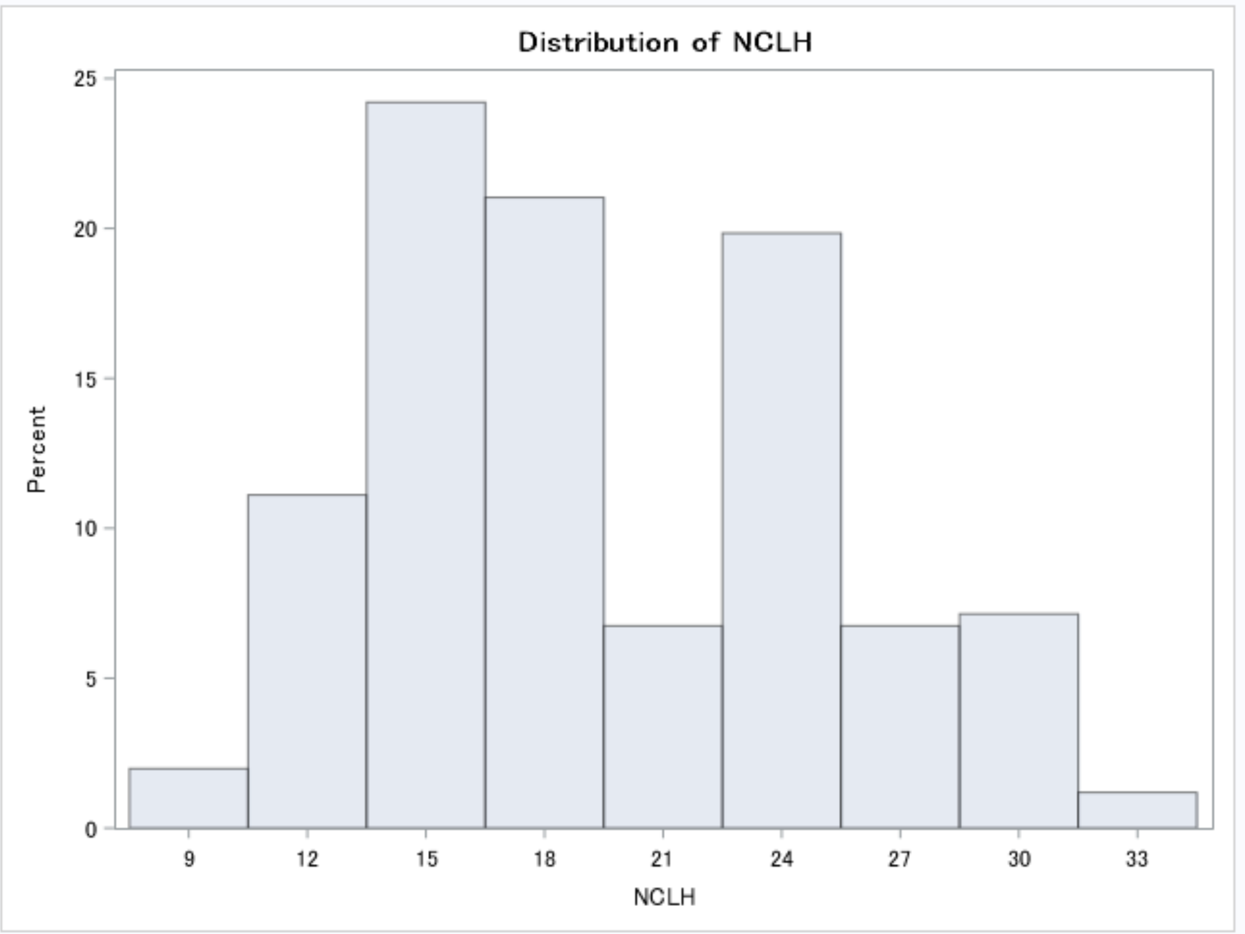


Figure 6: Distribution of NCLH

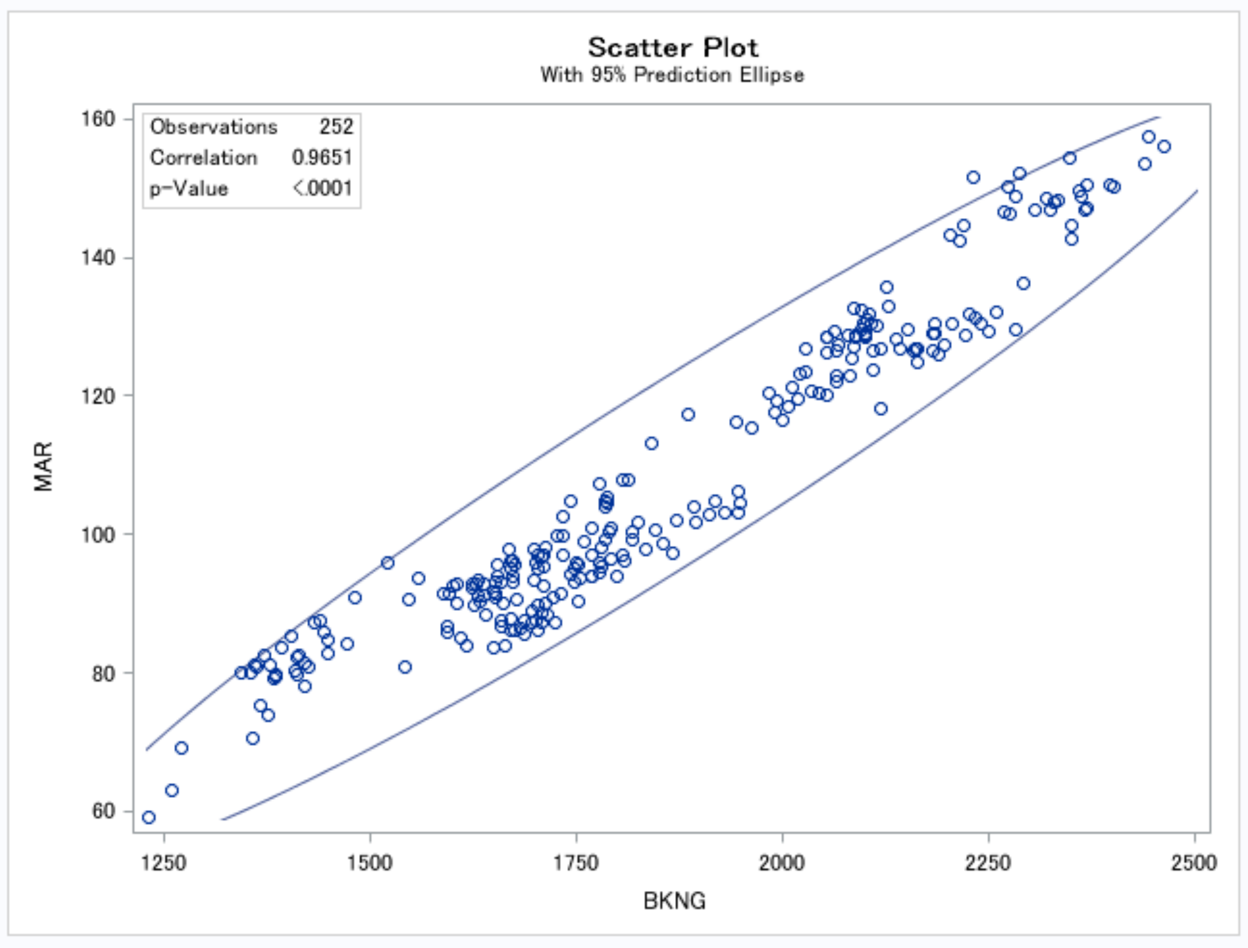


Figure 7 – BKNG vs MAR scatter plot

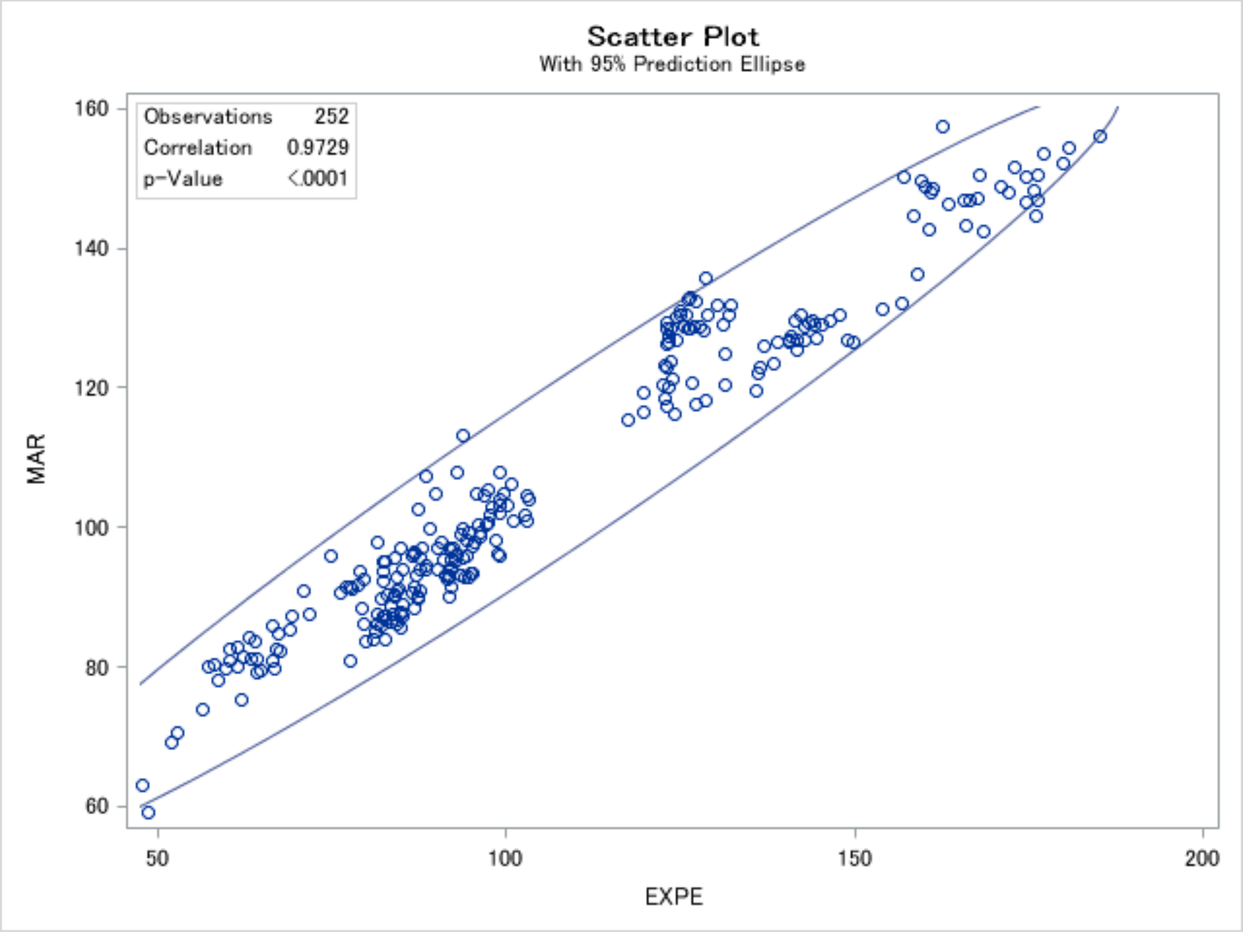


Figure 8: EXPE vs MAR scatter plot

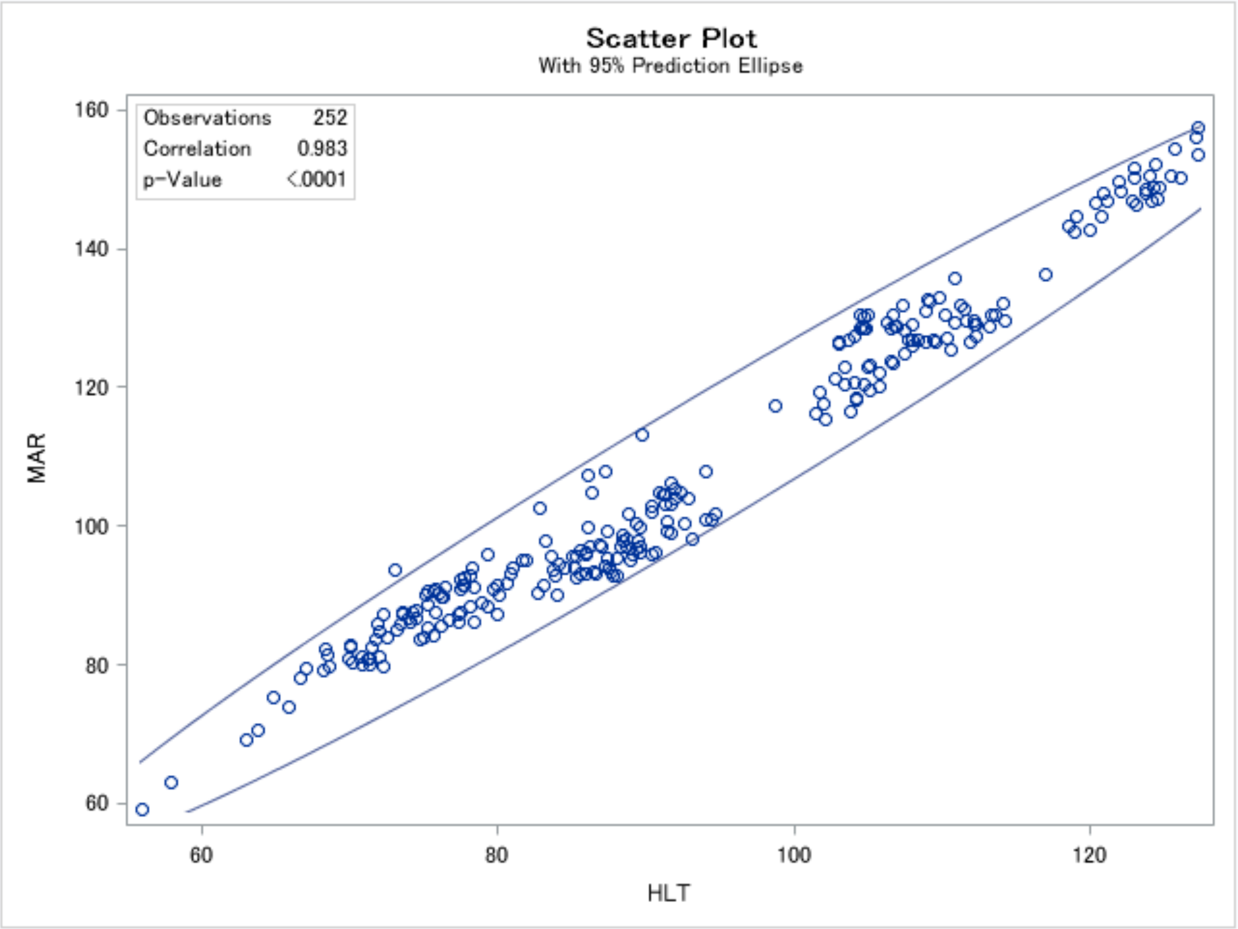


Figure 9: HLT vs MAR scatter plot

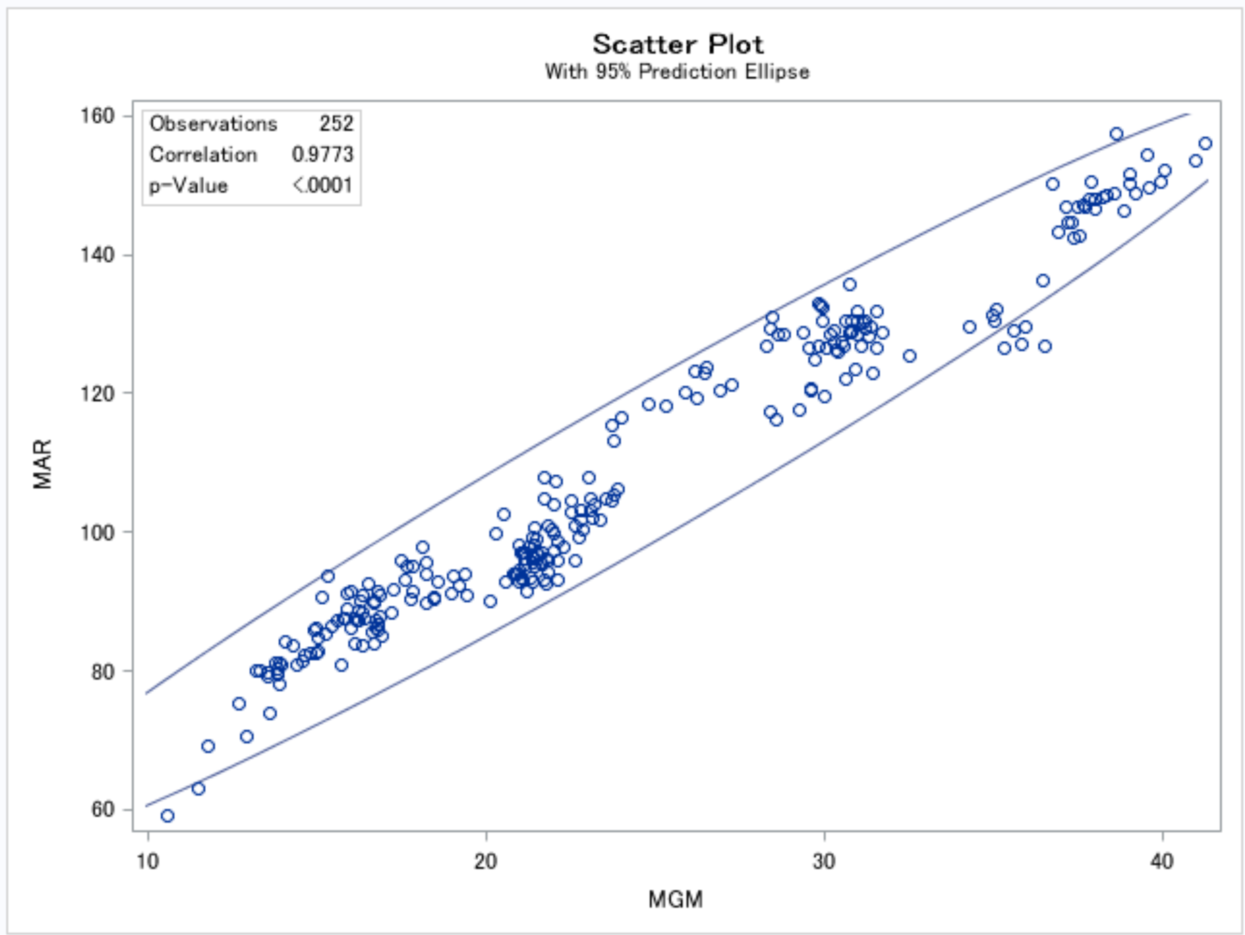


Figure 10: MGM vs MAR scatter plot

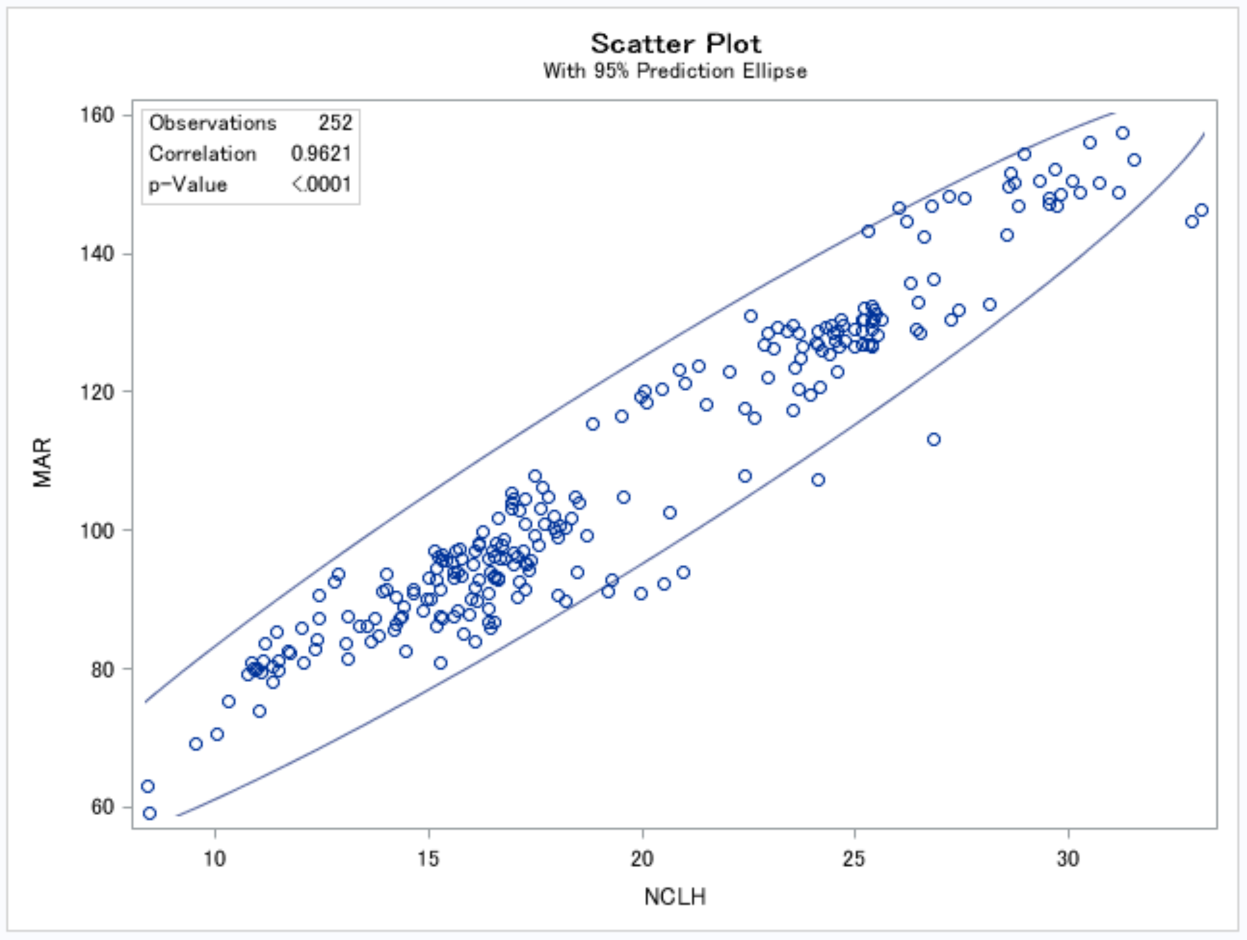


Figure 11: NCLH vs MAR scatter plot

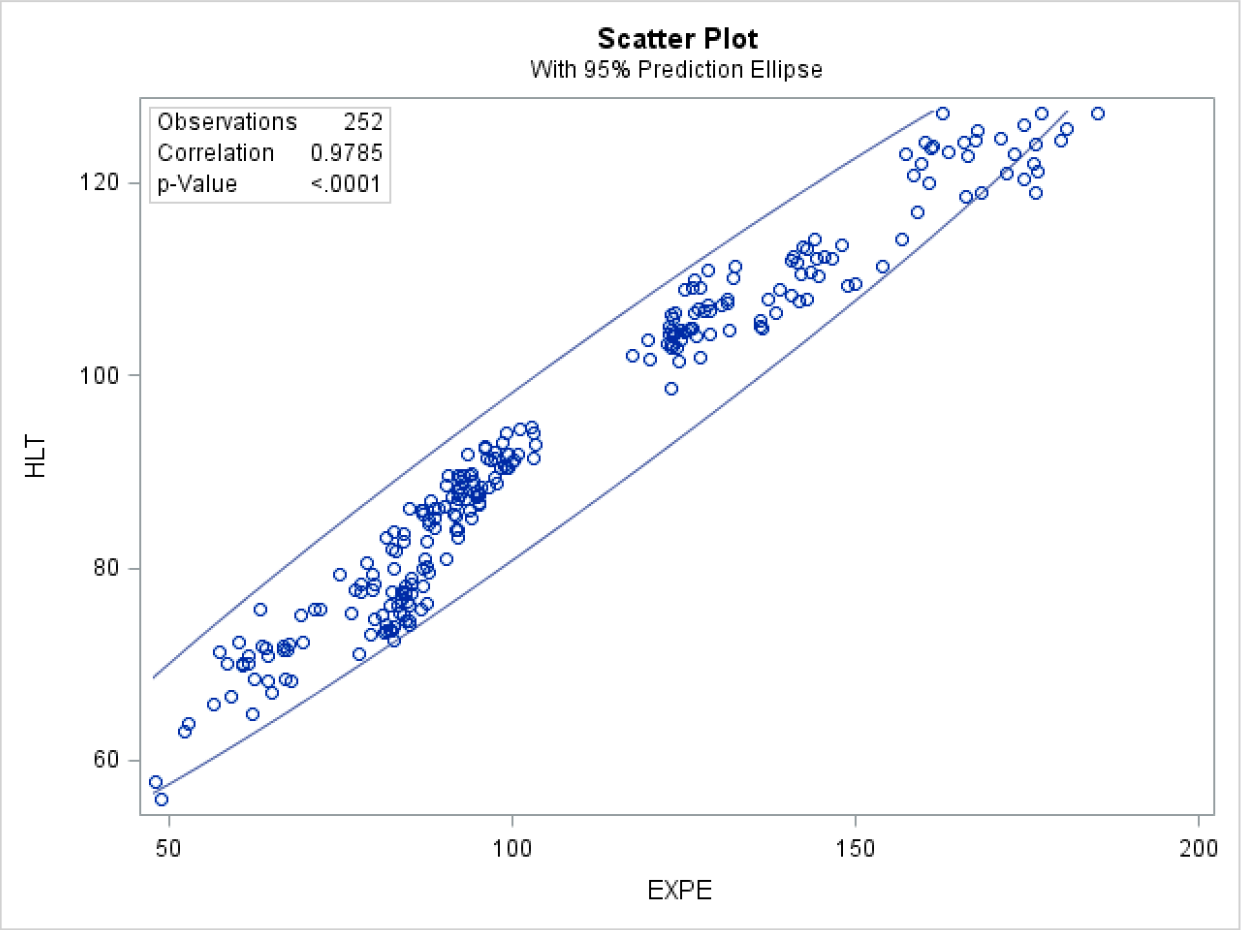
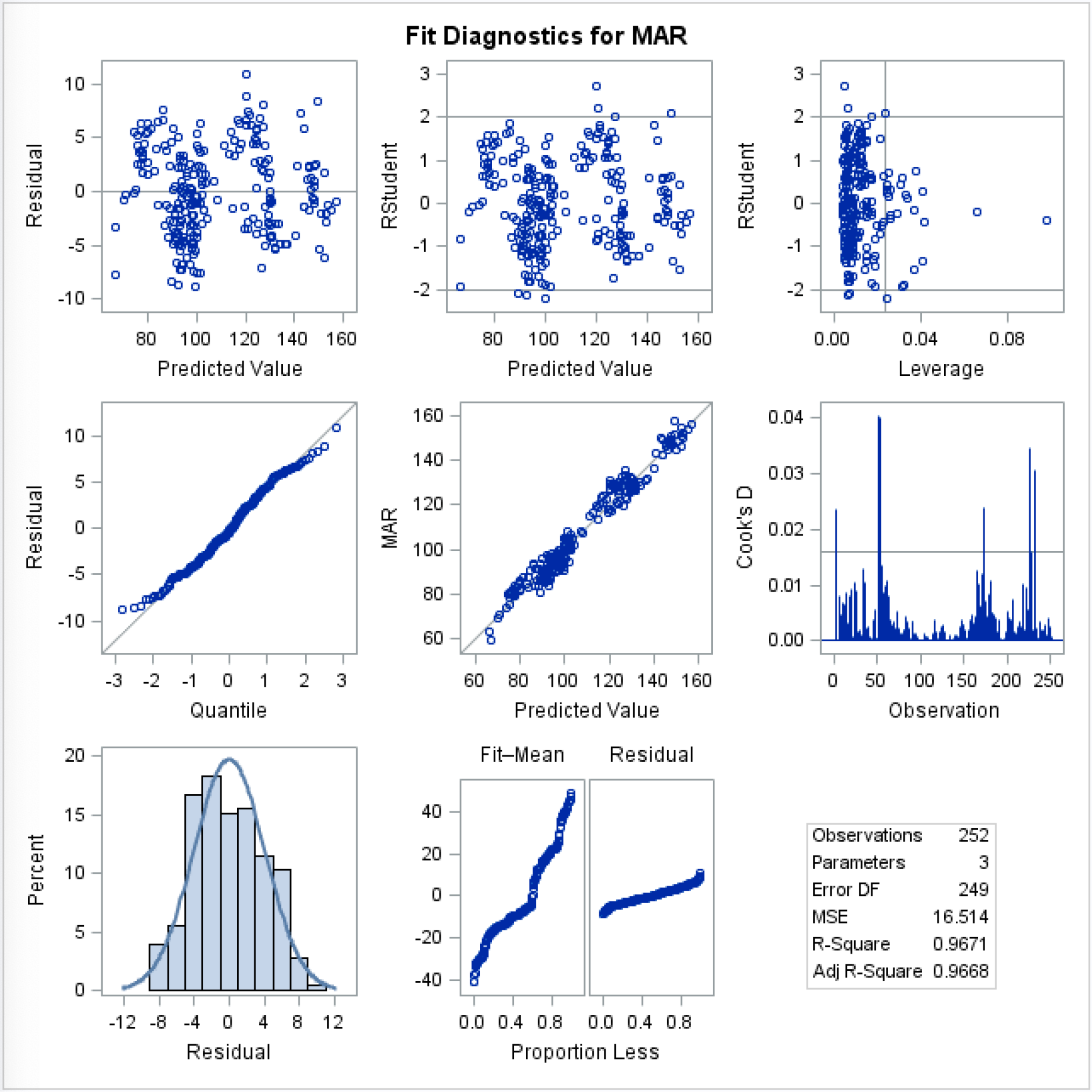


Figure 12: EXPE vs. HLT plot



```
Observations        252
Parameters            3
Error DF            249
MSE              16.514
R-Square         0.9671
Adj R-Square     0.9668
```

Figure 13: Residuals for final model