Data Description and Background

A university medical center's urology group was interested in the association between prostate specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical prostectomies. Each line of data set provides information on 8 other variables for each person.

Variable Name	Variable Description	Information
cavol	Cancer Volume	Estimate of prostate cancer volume (cc)
weight	Weight	Prostate weight (gm)
age	Age	Age of patient (years)
bph	Benign Prostatic Hyperplasia	Amount of benign prostatic hyperplasia (cm2) hyperplasia
svi	Seminal Vesicle Invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 if no
cp	Capsular Penetration	Degree of capsular penetration (cm)
gleason	Gleason Score	Pathologically determined grade of disease (6,7,8). Note, a higher Gleason score indicates worse prognosis.
psa	PSA Level	Serum prostate-specific antigen level (mg/ml)

PSA is commonly used as a screening mechanism for detecting prostate cancer. However, to be an efficient screening tool it is important that we understand how PSA levels relate to factors that may determine prognosis and outcome.

The PSA test measures the blood level of prostate-specific antigen, an enzyme produced by the prostate. PSA levels under 4 ng/mL (nanograms per milliliter) are generally considered normal, while levels over 4 ng/mL are considered abnormal (although in men over 65 levels upto 6.5 ng/mL may be acceptable, depending upon each laboratory's reference ranges). PSA levels between 4 and 10 ng/mL indicate a risk of prostate cancer higher than normal, but the risk does not seem to rise within this six- point range. When the PSA level is above 10 ng/mL, the association with cancer becomes stronger. However, PSA is not a perfect test. Some men with prostate cancer do not have an elevated PSA, and most men with an elevated PSA do not have prostate cancer. PSA levels can change for many reasons other than cancer. Two common causes of high PSA levels are enlargement of the prostate (benign prostatic hypertrophy (BPH)) and infection in the prostate (prostatitis).

Some of the variable names may look unfamiliar to you - please use resources on the web if you feel unsure as to what these variables measure. The section above is based on excerpts from Wikipedia.org, and you can also find variable definitions at:

http://www.prostate-cancer.org/resource/glossary.html

For example, a large tumor may invade surrounding tissue and penetrate the wall of the prostate (variable svi and cp). Also, benign hyperplasia is associated with higher PSA levels, but is non-cancerous (variable bph).

The goal of the analysis is to develop a model for PSA to be used for inferential purposes. Your model should be parsimonious, that is a model that balances both explanatory power with simplicity. To this end, you may employ any of the methods learned in class. Write up a report (5-7 pages) describing how you obtained this model. Below is a list of things to address in the report.

- Are all the assumptions needed to fit the model satisfied? Do any transformations need to be applied to the response and/or explanatory variables in order to correct for any model deviations?

- Are there any outliers in the dataset? Are they adversely affecting the estimates obtained using the least squares method?

- Recall that the goal of model building is not to build the model that best fits your particular dataset, but rather a model that can generalize. Consequently, what is the method you will employ to select a model? How many variables will you use? How did you model them (aka via polynomial terms, interactions or transformations)?

You must properly justify and support your methods in the report with the appropriate graph and diagnostics. Once you decide upon a model, perform the usual diagnostics to ensure that the necessary assumptions are satisfied.

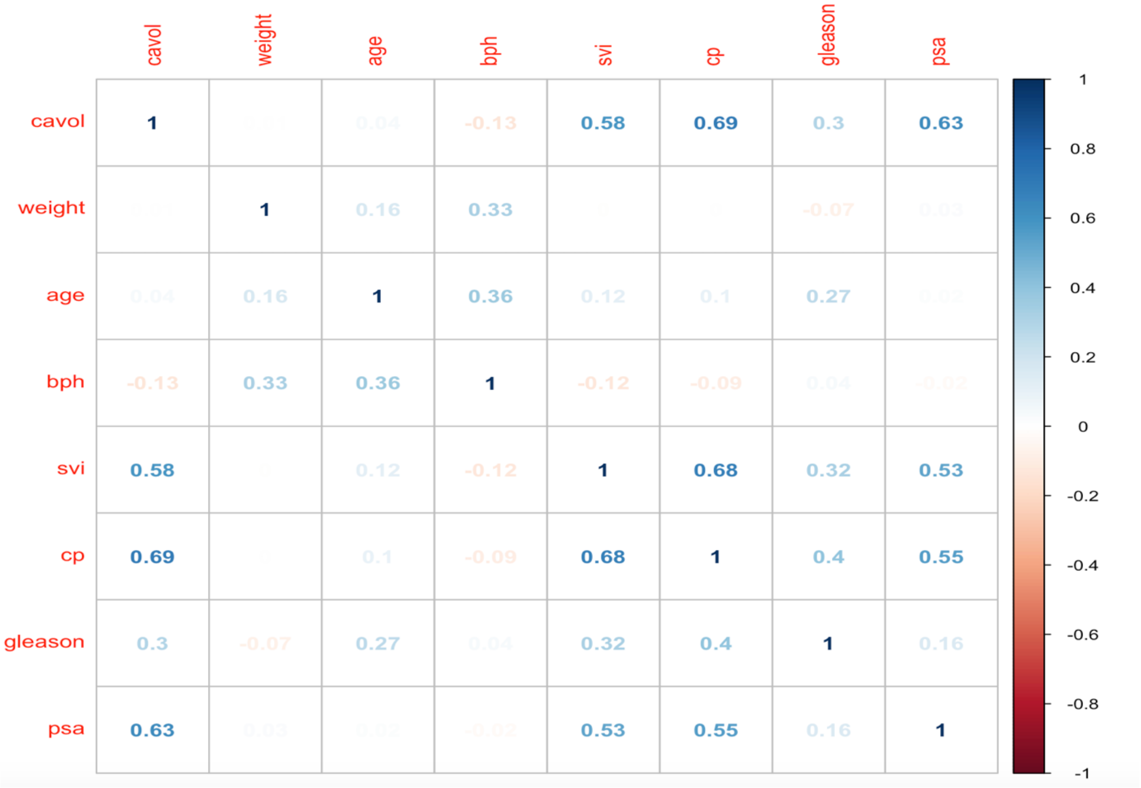
<u>Note:</u>

1) The entire analysis has been carried out in R.
2) Zoom to view the graphs attached in the documents properly.

Analysis

The primary goal of this analysis is to develop a model for PSA to be used for inferential purpose. For input dataset, we want to infer how the output is generated as a function of the data. The different steps employed for building the model are checking basic assumptions, selecting transformations as needed, handling outliers and model verification.

Correlation coefficient matrix between different combinations of input attributes is depicted below:



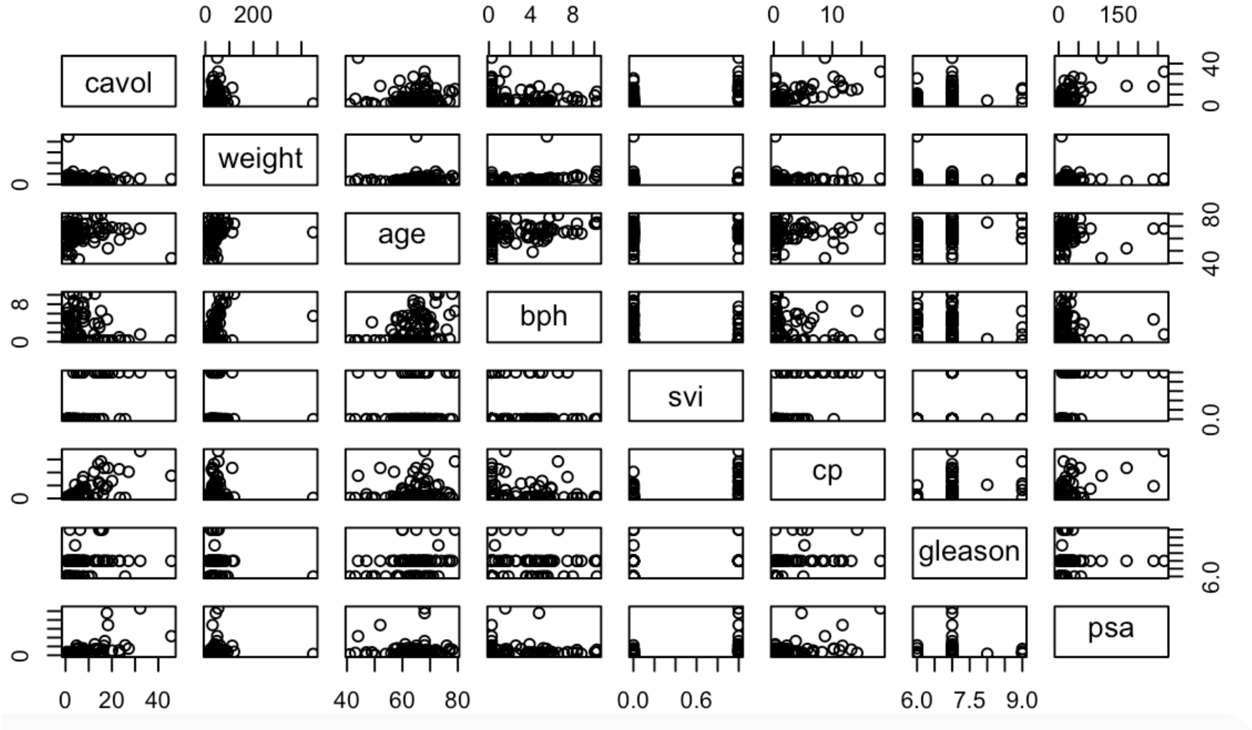
	cavol	weight	age	bph	svi	cp	gleason	psa
cavol	1	0.01	0.04	-0.13	0.58	0.69	0.3	0.63
weight	0.01	1	0.16	0.33	0	0	-0.07	0.03
age	0.04	0.16	1	0.36	0.12	0.1	0.27	0.02
bph	-0.13	0.33	0.36	1	-0.12	-0.09	0.04	-0.02
svi	0.58	0	0.12	-0.12	1	0.68	0.32	0.53
cp	0.69	0	0.1	-0.09	0.68	1	0.4	0.55
gleason	0.3	-0.07	0.27	0.04	0.32	0.4	1	0.16
psa	0.63	0.03	0.02	-0.02	0.53	0.55	0.16	1

From the plot above, it can be seen that strong positive correlation exists between the following pairs of input attributes:

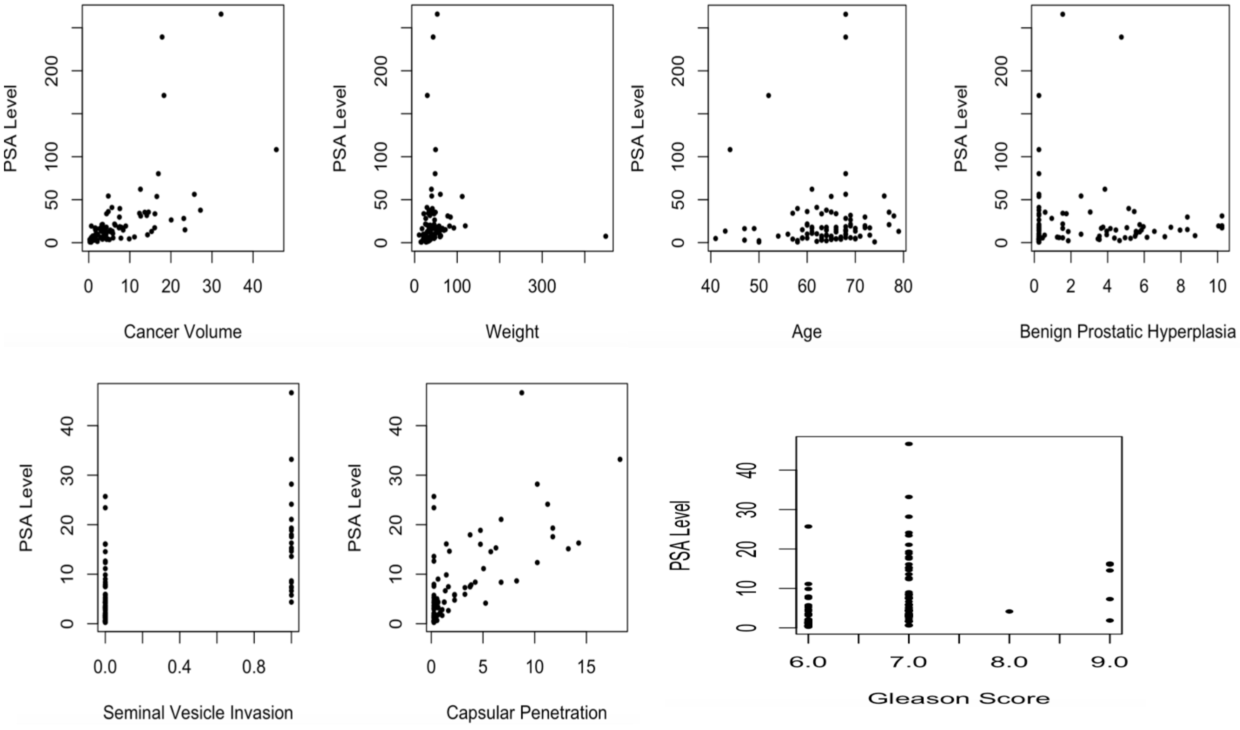
- cavol – cp
- cavol – svi
- cavol – psa
- svi – cp
- svi – psa
- cp - psa

There was **<u>no multicollinearity</u>** observed between attribute pairs based on correlation coefficients and VIF calculated for explanatory variables.

Pairwise scatter plot between different combinations of input attributes is depicted below:



Scatter plots between response variable (PSA) and individual explanatory variables (Cancer Volume - cavol, weight, age, Benign Prostatic Hyperplasia – bph, Seminal Vesicle Invasion - svi, Capsular Penetration - cp, Gleason Score - gleason) are depicted below:



From the plots above, it can be inferred that there appears to be a somewhat linear relationship between response variable and explanatory variables **cavol, weight, cp**. This needs to be explored further.

Summary statistics of simple linear models generated based on previous graphs depicted below:

psa ~ cavol	**psa ~ weight**
Residuals: Min 1Q Median 3Q Max -61.921 -8.988 -1.607 3.196 180.371 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 1.0855 4.3597 0.249 0.804 cavol 3.2359 0.4146 7.804 7.77e-12 *** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 32.04 on 95 degrees of freedom Multiple R-squared: 0.3906, Adjusted R-squared: 0.3842 F-statistic: 60.9 on 1 and 95 DF, p-value: 7.771e-12	Residuals: Min 1Q Median 3Q Max -25.984 -17.900 -10.149 -2.476 241.924 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 22.64780 5.89755 3.840 0.000222 *** weight 0.02401 0.09180 0.262 0.794246 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 41.02 on 95 degrees of freedom Multiple R-squared: 0.0007195, Adjusted R-squared: -0.009799 F-statistic: 0.0684 on 1 and 95 DF, p-value: 0.7942

psa ~ age	**psa ~ bph**
Residuals: Min 1Q Median 3Q Max -23.807 -17.873 -10.630 -2.954 241.736 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 17.9599 36.1666 0.497 0.621 age 0.0905 0.5625 0.161 0.873 Residual standard error: 41.03 on 95 degrees of freedom Multiple R-squared: 0.0002724, Adjusted R-squared: -0.01025 F-statistic: 0.02588 on 1 and 95 DF, p-value: 0.8725	Residuals: Min 1Q Median 3Q Max -23.761 -18.060 -10.024 -2.062 241.804 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 24.4807 5.6196 4.356 3.34e-05 *** bph -0.2802 1.4260 -0.196 0.845 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 41.03 on 95 degrees of freedom Multiple R-squared: 0.0004061, Adjusted R-squared: -0.01012 F-statistic: 0.0386 on 1 and 95 DF, p-value: 0.8447

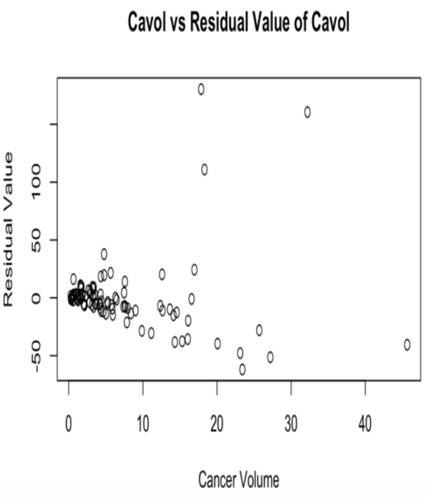
psa ~ svi	**psa ~ cp**
Residuals: Min 1Q Median 3Q Max -55.424 -9.607 -4.907 4.093 201.276 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 12.457 3.996 3.117 0.00242 ** svi 52.117 8.589 6.068 2.64e-08 *** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 34.84 on 95 degrees of freedom Multiple R-squared: 0.2793, Adjusted R-squared: 0.2717	Residuals: Min 1Q Median 3Q Max -82.910 -9.457 -4.802 5.443 201.005 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 9.3877 4.1173 2.280 0.0248 * cp 6.0752 0.9379 6.478 4.1e-09 *** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 34.18 on 95 degrees of freedom Multiple R-squared: 0.3064, Adjusted R-squared: 0.2991

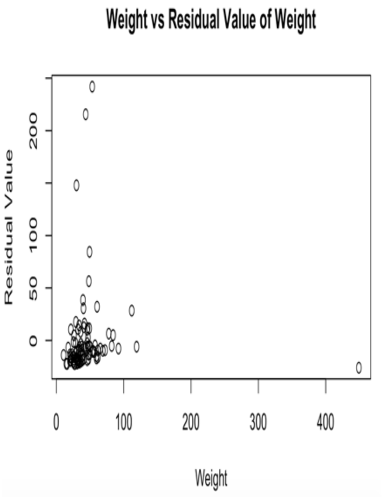
F-statistic: 36.82 on 1 and 95 DF, p-value: 2.636e-08	F-statistic: 41.96 on 1 and 95 DF, p-value: 4.099e-09

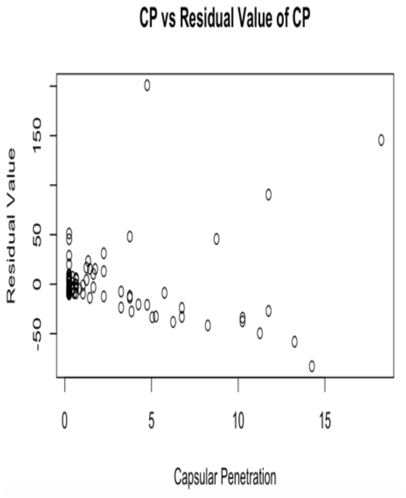
psa ~ gleason

```
Residuals:
    Min      1Q  Median      3Q     Max
-33.900 -15.204 -10.356   0.144 239.896

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -36.682     38.891  -0.943    0.348
gleason        8.948      5.727   1.562    0.122

Residual standard error: 40.52 on 95 degrees of freedom
Multiple R-squared:  0.02505,	Adjusted R-squared:  0.01479
F-statistic: 2.441 on 1 and 95 DF,  p-value: 0.1215
```

From the plots, it appears that some of the non-categorical explanatory variables **(cavol, weight, cp)** has a somewhat linear relationship with the response variable. It is also observed from the summary statistics that the **p-value is significant for variables cavol, svi, cp, gleason**. The residual plots for these variables indicate variance or presence of **heteroscedasticity** which needs to be corrected by applying suitable transformation. Residual plots are depicted below:



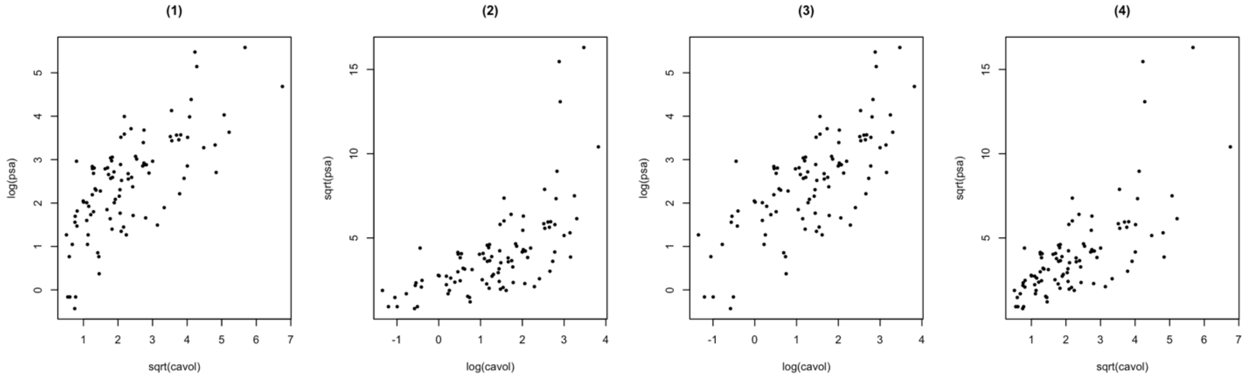




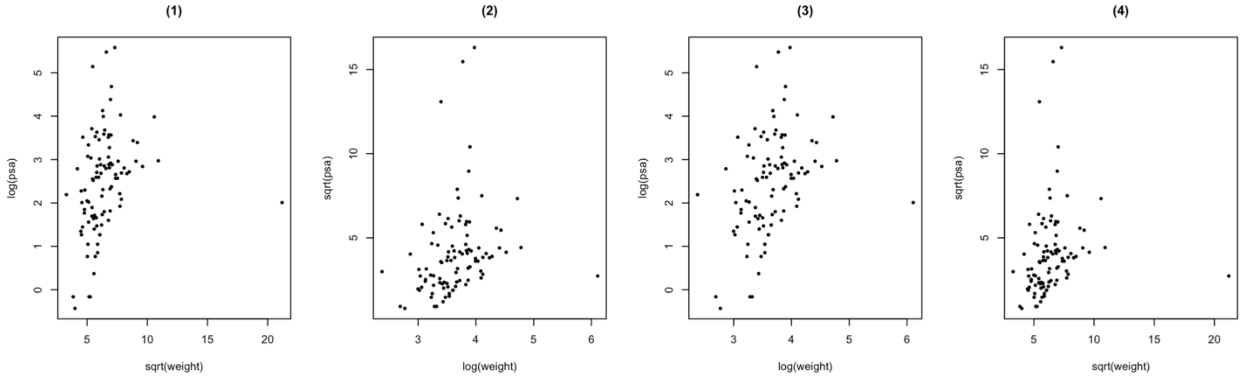
Note: Variable svi is a categorical variable and gleason score consists of a fixed set of values (6,7,8,9).

To address heteroscedasticity, different combinations of transformations **(natural logarithm and square root)** are carried out between explanatory (cavol, weight, cp) and response variable (psa). Transformations are not carried for categorical variables (svi) and fixed value variables (gleason). The plots are depicted below:

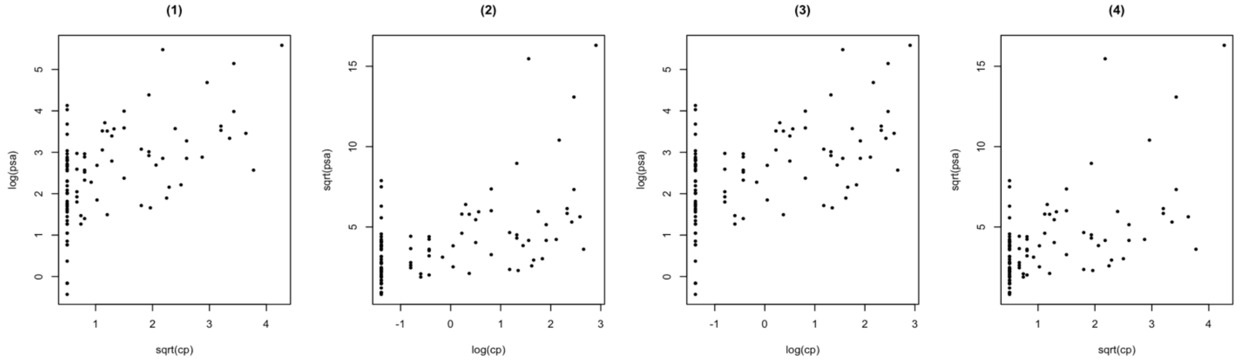
cavol



weight



cp



From the plots it can be seen that by applying log transformation of both explanatory and response variable, heteroscedasticity is corrected to an extent for **cavol and weight** explanatory variables. For **cp** variable, it was attempted to analyze if there is any **polynomial relationship** with response variable. The R code block for creating model summary of polynomial transformations is depicted below:

```
## CP Polynomial Transformation ##

summ_cp_1 = summary(lm(log(psa)~cp))
summ_cp_1  ## Adjusted R-squared:  0.2559, p-value: 7.536e-08

summ_cp_2 = summary(lm(log(psa)~cp + I(cp^2)))
summ_cp_2  ## Adjusted R-squared:  0.2561 , p-value: 3.405e-07

summ_cp_3 = summary(lm(log(psa)~cp + I(cp^2) + I(cp^3)))
summ_cp_3  ## Adjusted R-squared:  0.2804 , p-value: 2.26e-07

summ_cp_4 = summary(lm(log(psa)~cp + I(cp^2) + I(cp^3) + I(cp^4)))
summ_cp_4  ## Adjusted R-squared:  0.2762 , p-value: 7.421e-07
```

From the summary above, it can be observed that Adjusted R-squared value is best for cubic polynomial. Moreover, p-values are significant for all polynomial models. So, **relationship between explanatory variable cp and response variable psa is cubic polynomial.**

From the scatter plot created before for explanatory variables age and bph, it can be clearly seen that the relationship is not linear, monotonic or simple. Consequently, we didn’t get a significant p-value for these variables. It was attempted to analyze if there is any **polynomial relationship** with response variable. The R code block creating model summary depicted below:

```
## Age Polynomial Transformation ##

summ_age_1 = summary(lm(log(psa)~age))
summ_age_1  ## Adjusted R-squared:  0.01854, p-value: 0.09677

summ_age_2 = summary(lm(log(psa)~age + I(age^2)))
summ_age_2  ## Adjusted R-squared:  0.01941 , p-value: 0.148

summ_age_3 = summary(lm(log(psa)~age + I(age^2) + I(age^3)))
summ_age_3  ## Adjusted R-squared:  0.0144 , p-value: 0.2285

summ_age_4 = summary(lm(log(psa)~age + I(age^2) + I(age^3) + I(age^4)))
summ_age_4  ## Adjusted R-squared:  0.003854 , p-value: 0.3648
```

```
## Bph Polynomial Transformation ##

summ_bph_1 = summary(lm(log(psa)~bph))
summ_bph_1  ## Adjusted R-squared:  0.01345, p-value: 0.132

summ_bph_2 = summary(lm(log(psa)~bph + I(bph^2)))
summ_bph_2  ## Adjusted R-squared:  0.004058 , p-value: 0.3071

summ_bph_3 = summary(lm(log(psa)~bph + I(bph^2) + I(bph^3)))
summ_bph_3  ## Adjusted R-squared:  0.008516 , p-value: 0.2876

summ_bph_4 = summary(lm(log(psa)~bph + I(bph^2) + I(bph^3) + I(bph^4)))
summ_bph_4  ## Adjusted R-squared:  0.001603 , p-value: 0.3917
```

From the model summaries above, it can be seen that **Quadratic polynomial has highest Adjusted R-squared value for both age and bph**. It is also observed that, **p-value is not significant for any of the polynomial transformations (quadratic, cubic, quartic) for both age and bph explanatory variables.** Hence, these variables can be excluded from any kind of model building going forward.

So, for building **multiple regression model (MLM)**, the following variables will be used based on analysis till now:

Explanatory Variables:

- cavol (log transformed)
- weight (log transformed)
- cp (cubic polynomial transformation)
- svi (categorical variable)
- gleason

Response Variable:

- psa (log transformed)

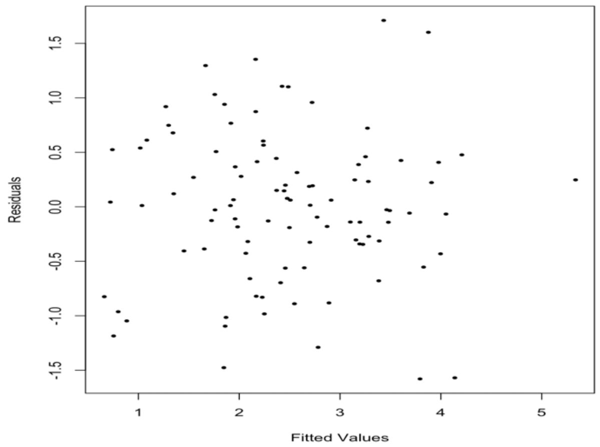
Model Output Summary:

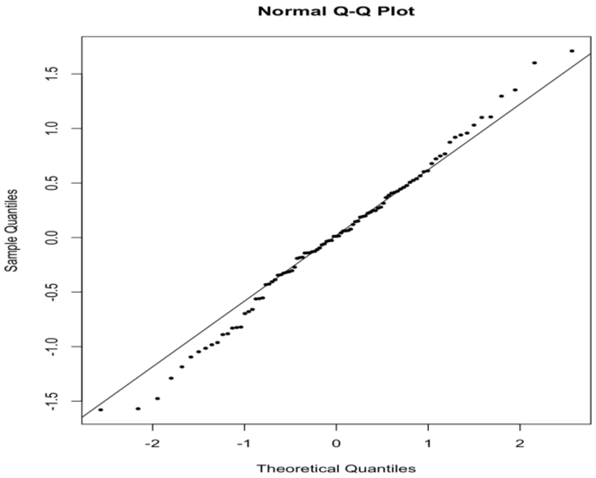
```
Call:
lm(formula = log(psa) ~ log(cavol) + log(weight) + svi + cp +
I(cp^2) + I(cp^3) + gleason)

Residuals:
Min       1Q   Median       3Q      Max
-1.57937 -0.38676  0.01144  0.42526  1.71022

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.1100854  0.9618881  -1.154 0.251561
log(cavol)   0.5671679  0.0867323   6.539 3.79e-09 ***
log(weight)  0.5102829  0.1497771   3.407 0.000988 ***
svi          0.7929598  0.2476377   3.202 0.001893 **
cp          -0.0211482  0.1348878  -0.157 0.875770
I(cp^2)     -0.0122182  0.0203440  -0.601 0.549648
I(cp^3)      0.0008544  0.0008256   1.035 0.303539
gleason      0.1310454  0.1186742   1.104 0.272463
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.711 on 89 degrees of freedom
Multiple R-squared:  0.6483,           Adjusted R-squared:  0.6206
F-statistic: 23.44 on 7 and 89 DF,  p-value: < 2.2e-16
```

Conclusion: Following inferences can be drawn from the MLM model summary above:

- Adjusted R-squared value indicates that the final transformed model is able to explain **62% deviation** in response.
- The F-statistic and p-value indicates that the overall model is statistically significant.

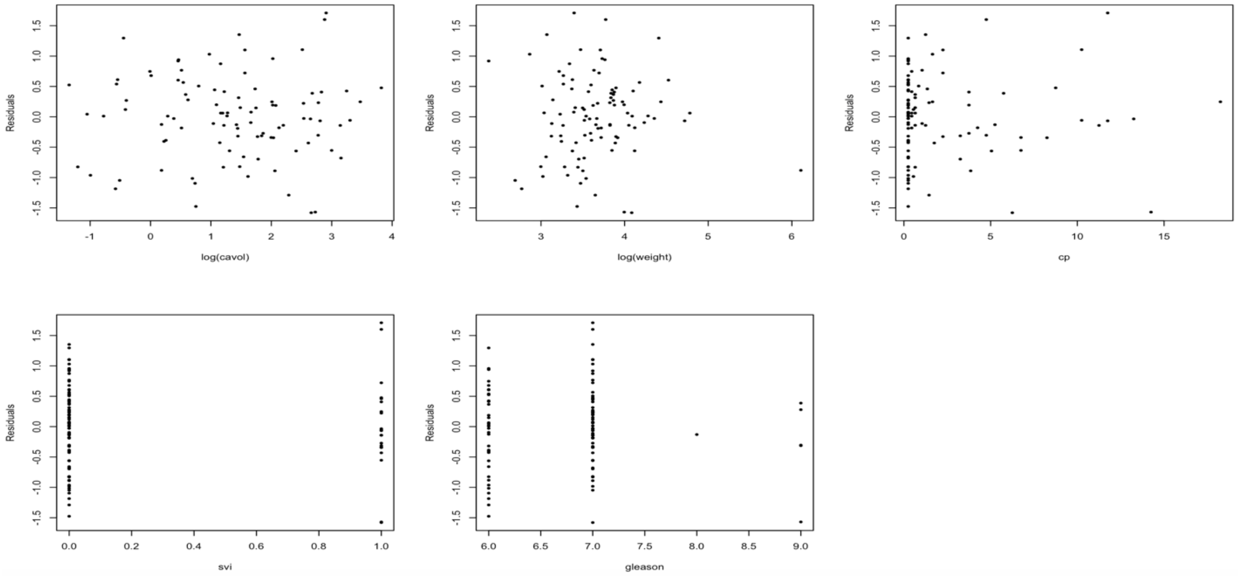
The residual plot and the QQ Quantile plot for the transformed model is depicted below:





Residual plot indicates that heteroscedasticity is corrected in transformed model and appear to be roughly distributed around 0. QQ-plot indicates that observations lie 'roughly' on a straight line and its same to assume that the residuals are arising from a normal distribution after transforming variables.

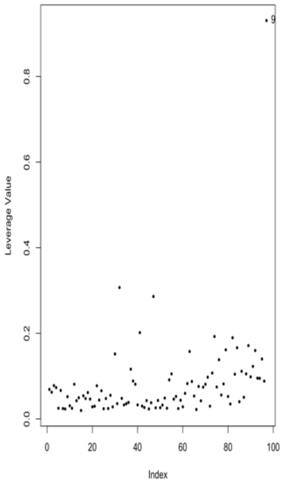
Residual plots between individual transformed variables and residuals are depicted below. Residuals appear to be roughly evenly distributed around 0 and looks normal.

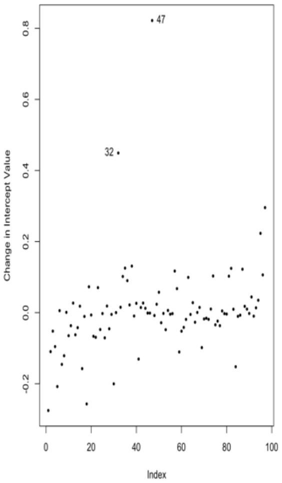


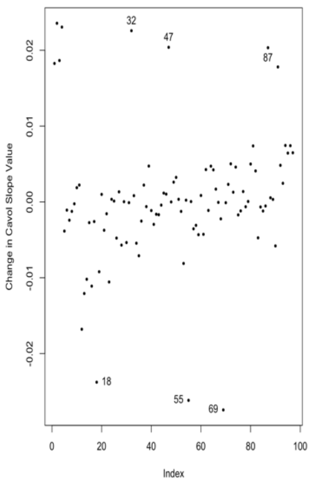
Outlier Detection: The final transformed model was investigated to understand whether one or more observations are outlying with respect to their X values and therefore might be excessively influencing the regression results. The following methods are employed to detect outliers:

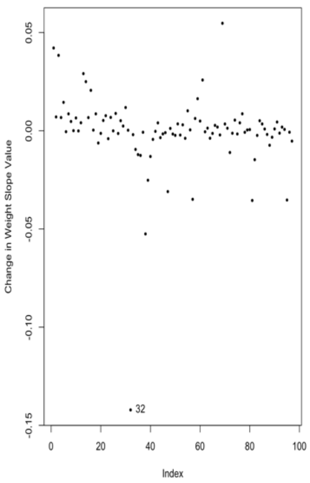
- By calculating leverage
- By calculating cooks' distance
- By calculating change in fitted values (dfbeta)
- By using influence function to assess whether an observation is an influential observation

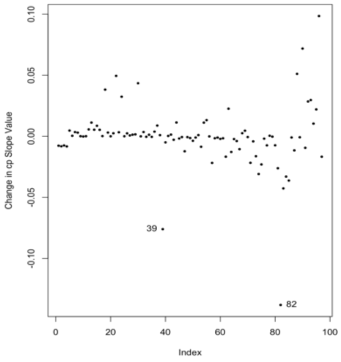
All the generated plots for outlier detection are depicted below:

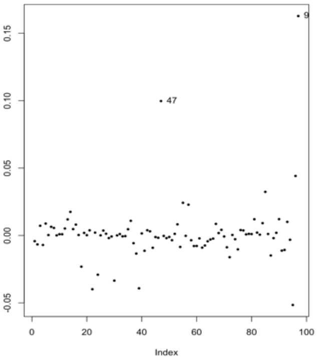


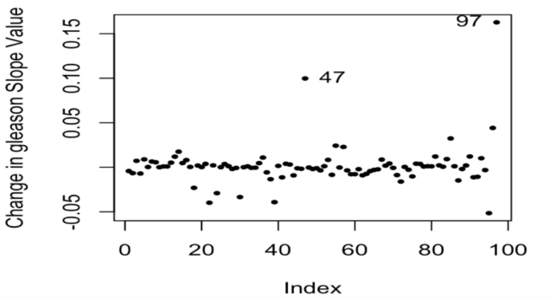


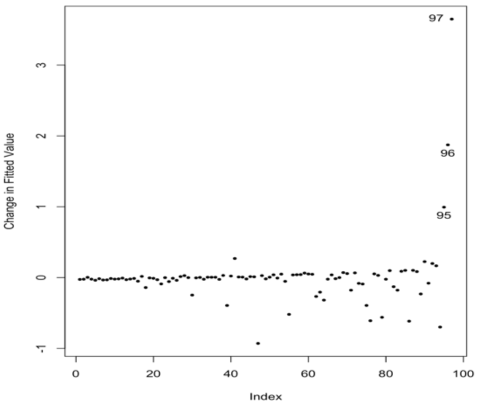


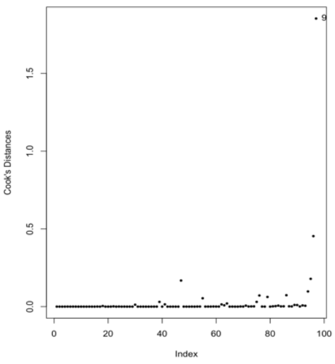


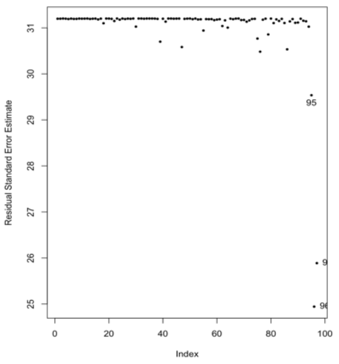












From the above plots, it can be seen that a certain set of observations are repeated. These observations were further analyzed to check for abnormality. The R code snippet is depicted below:

```
## Analyzing observations 32, 47, 55, 95, 96, 97 ##

prostate[32,]   ## Weight is 449.25. Invalid data.
prostate[47,]   ## Nothing abnormal
prostate[55,]   ## PSA is high because of high cavol. Nothing abnormal
prostate[95,]   ## Nothing Abnormal. Valid data.
prostate[96,]   ## Nothing Abnormal. Valid data.
prostate[97,]   ## Nothing Abnormal. Valid data.
```

Based on above summary, **observation 32 was removed** from the input dataset because it appeared to be erroneous. After removing outlying observations (32), MLM model was rebuilt with input dataset. The summary of the revised model is depicted below:

```
Call:
lm(formula = log(psa[-32]) ~ log(cavol[-32]) + log(weight[-32]) +
    svi[-32] + cp[-32] + I(cp[-32]^2) + I(cp[-32]^3) + gleason[-32])

Residuals:
     Min       1Q   Median       3Q      Max
-1.62549 -0.38566 -0.02574  0.40177  1.76031

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.5592612  1.0010017  -1.558 0.122893
log(cavol[-32])   0.5445917  0.0874331   6.229 1.57e-08 ***
log(weight[-32])  0.6524294  0.1763405   3.700 0.000375 ***
svi[-32]          0.7894865  0.2459254   3.210 0.001852 **
cp[-32]          -0.0222023  0.1339511  -0.166 0.868735
I(cp[-32]^2)     -0.0117358  0.0202050  -0.581 0.562837
I(cp[-32]^3)      0.0008292  0.0008201   1.011 0.314688
gleason[-32]      0.1270515  0.1178785   1.078 0.284061
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.706 on 88 degrees of freedom
Multiple R-squared:  0.6565,  Adjusted R-squared:  0.6292
F-statistic: 24.02 on 7 and 88 DF,  p-value: < 2.2e-16
```

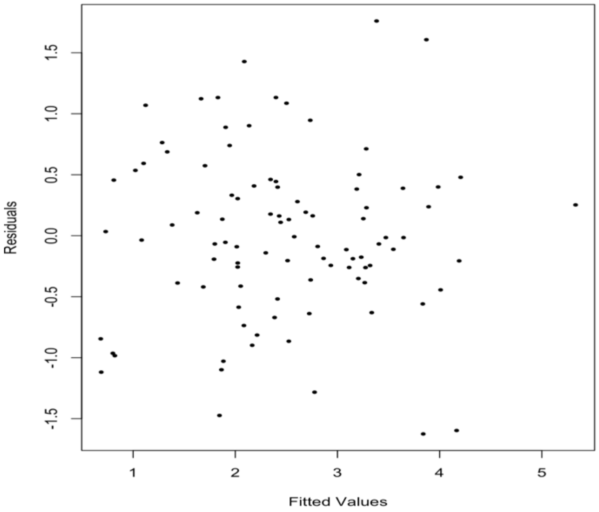
Conclusion:

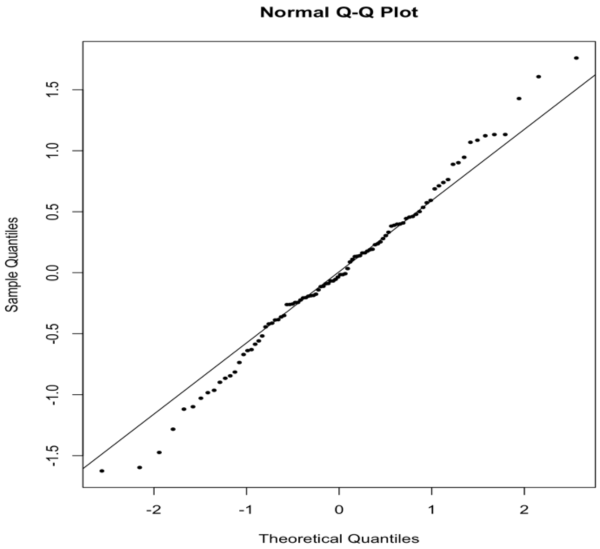
- After removing outlier, the model performance improved slightly. Adjusted R-squared **increased from 0.6206 to 0.6292**. So, the new revised model was able to explain **~ 62.92%** deviation of response variable.
- F-statistic also improved from 23.44 to 24.02.
- Overall p-value also remained significant.

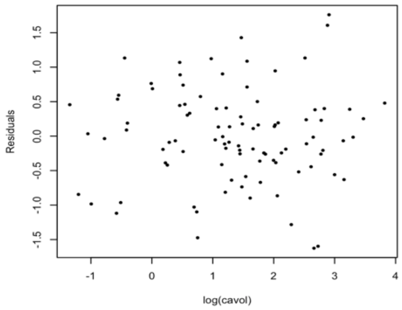
<u>Further Analysis:</u>

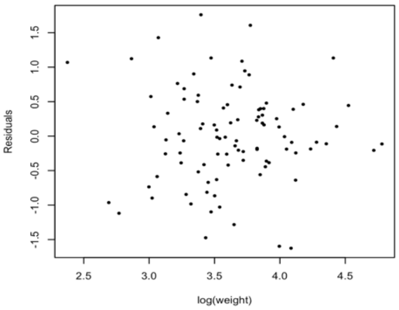
- AIC and BIC indicators were used to estimate the quality of all models generated. Lower scores indicate better performance. The final transformed model excluding outlier gave best scores i.e. **AIC Score = 215.2477 and BIC score = 238.3268.**
- Correlation coefficient between fitted values of psa vs response variable (psa) is **0.8102334** which indicates strong correlation.

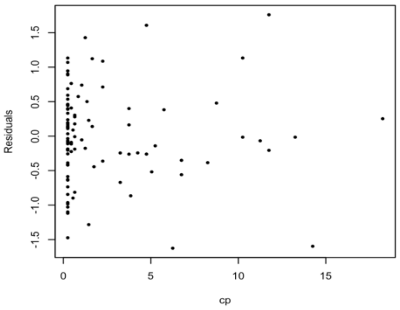
<u>Final Graphs:</u> All residual plots and the QQ Quantile plot for the final transformed model after removing outlier are depicted below. Residual plots look consistent with Homoscedasticity assumption and QQ plot also looks normal.

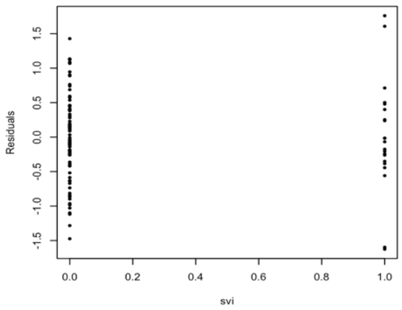


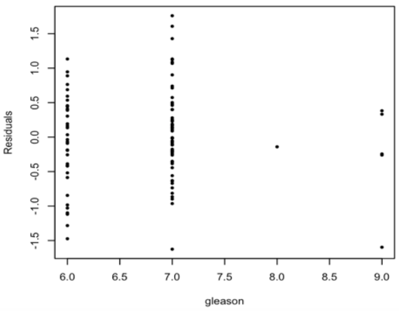












Final Conclusion:

- **Cavol, weight, svi, cp and gleason** were the explanatory variables which had a significant relationship with response variable psa. This was determined on the basis of p-values obtained from individual models.

- Attributes **bph and age** didn't show any significance even after transformation. Hence, they were excluded from being part of final aggregate multiple linear model.

- Final transformed model was chosen based on certain transformation. Explanatory variable cavol and weight was log transformed and cubic polynomial transformation was chosen for cp. Response variable psa was also log transformed. This transformed model gave the best Adjusted R-Squared estimates i.e. **~ 62.9%**. The initial MLR model built using all untransformed variables gave an Adjusted R-Squared value of **~ 41%**.

- Different methods were employed to detect presence of outliers. One observation was removed which resulted in marginally improved performance of the final model.

- Correlation coefficient between fitted values of psa vs response variable (psa) was **0.8102334** which indicated strong correlation.

- For making prediction, inverse transformation needs to be done for transformed explanatory and response variable before the model can be used for prediction. Confidence and prediction intervals can also be derived from the same. The scope of this project was limited to building an inferential model. It can be extended to make predictions for new data points.

- More data is needed to improve final transformed model i.e. better Adjusted R-Squared value.