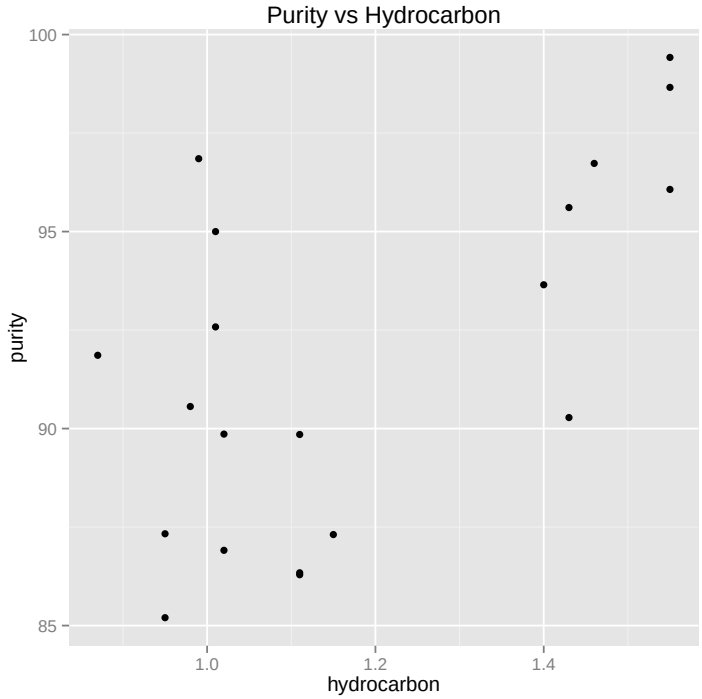
# Question 2

**A:** Plot Purity vs Hydrocarbon. Discuss what you see in relation to the SLR assumptions.

**Assumtion 1:** There does apear to be a linear relationship between purity and hydrocarbons. The expectation of residuals does apear to be approximetly zero. **Assumption 2:** The error or residual values do not have constant variances. As seen from the plot the hypothetical error variables are considerbly more variable for lower hydrocarbon values than for higher ones. The homoscedasticity assumption might be violated in this case. **Assumption 3:** The error/residuals do seem to be independent from one another. **Assumption 4:** Even though their variance isn't constant, the error values do seem to be normaly distributed.

```
> ## loading useful packages
> library(knitr);library(xtable);library(ggplot2)
> ## read in data
> data<-read.table("A1_data.txt",sep=" ",header=T)
> ## rename the variables
> names(data)<-c("purity","hydrocarbon")
> ## Plot graph
> plot_1<-qplot(hydrocarbon,purity,data=data,main="Purity vs Hydrocarbon")
> print(plot_1)
```



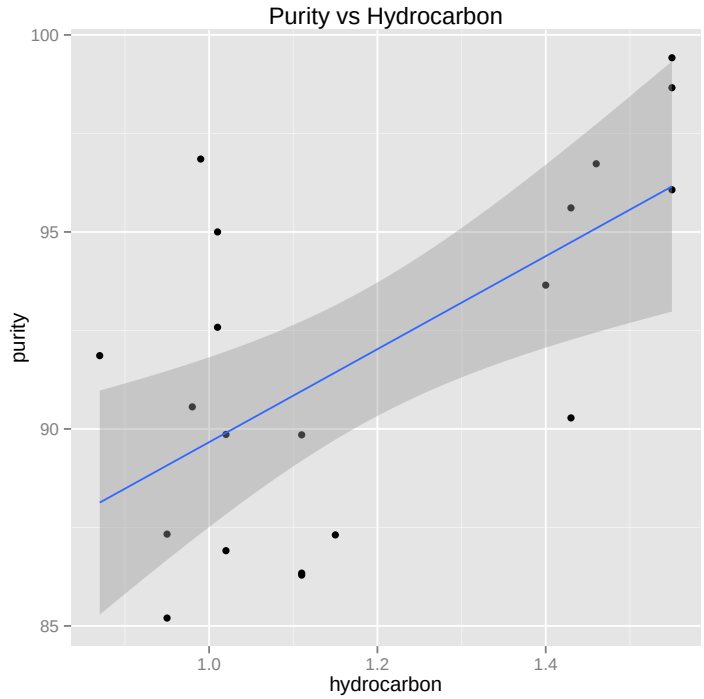
**B:** Fit a linear model and plot the fitted line to (A). Explain the model

```
> fit1<-lm(purity~hydrocarbon,data=data)
> print(xtable(summary(fit1)))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 77.8633 | 4.1989 | 18.54 | 0.0000 |
| hydrocarbon | 11.8010 | 3.4851 | 3.39 | 0.0033 |

The value of slope or the B1 estimator, 11.8010281931501 is interpreted as the expected change in purity given a 1 unit change of hydrocarbons. Similarly,the intercept value of 77.8632841616 or the B0 estimator represents the expected value of purity at 0 hydrocarbons, though this might not make any contextual sense in some cases.

```
> plot_1<-plot_1+geom_smooth(method = "lm")
> print(plot_1)
>
```



**Some Diagnostics**

In terms of the overall statistical significance of the model. We look at the F-test for the model using the Anova table.

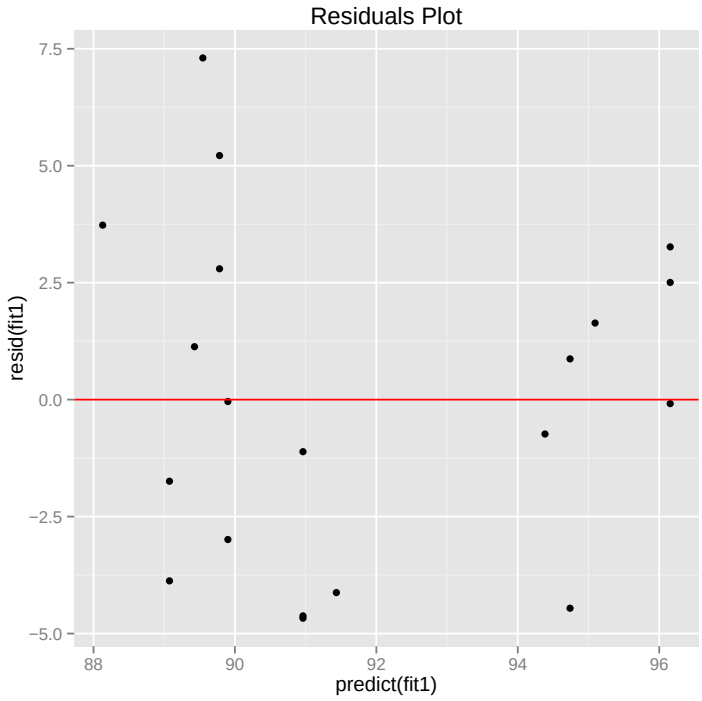
```
> print(xtable(anova(fit1)))
```

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| hydrocarbon | 1 | 148.31 | 148.31 | 11.47 | 0.0033 |
| Residuals | 18 | 232.83 | 12.94 | | |

As we can see the p-values for the corresponding F-test is quite low leading us to believe that the model does have some explanatory power. The R-squared values turns out to be 0.389122405808525 which is quite low. This result shows that this model is not particularly good at explaining the variance of the purity variable.

As mentioned earlier the constant variance assumption might be violated in our data set. In order to check for this the residual vs predicted plot is a usefull tool.

```
>  qplot(predict(fit1),resid(fit1))+ggtitle("Residuals Plot") +
+    geom_hline(yintercept = 0,colour="red")
```



As we can clearly see from the residuals plot the variance is does have a pattern. The residuals are definetly more variable for low predicted values, which supports our previous analysis.

To summarize, due to the low R-squared values and the apparent violation of the constant variance assumption this model is unlikely to have any significant explanatory or predicitive value.

**C:** Fit a 95 percent prediction and confidence interval for the purity level when the hydrocarbon percentage is equal to to 1.0. Explain the PI and CI.

```
> confidence_interval<-predict(fit1,newdata=data.frame(hydrocarbon=1),
+                              interval=("confidence"),level=0.95)
> prediction_interval<-predict(fit1,newdata=data.frame(hydrocarbon=1),
+                              interval=("prediction"),level=0.95)
> print(confidence_interval)

       fit      lwr      upr
1 89.66431 87.51017 91.81845

> print(prediction_interval)

       fit      lwr      upr
1 89.66431 81.80716 97.52146
```

A confidence interval expresses uncertainty about the expected value of y-values at a given x. A prediction interval expresses uncertainty surrounding the predicted y-value of a single sampled point with that value of x. In regards, to our experiment we say that the confidence interval is the interval in which we say that 95 percent of the sample means we observe will fall in this interval, given that the hydrocarbon value is equal to 1.0. The prediction interval is the interval in which we are 95 percent certain the values of the next observation will fall within this inteval, again given that the hydrocarbon value is zero. As such, the prediction interval is wider than the confidence interval.

**D:** Do the hypothesis test for B1=0 at 0.05 level of significance. Explain.

```
> print(xtable(summary(fit1)))
```

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 77.8633 | 4.1989 | 18.54 | 0.0000 |
| hydrocarbon | 11.8010 | 3.4851 | 3.39 | 0.0033 |

As we can see the table summary of coefficients automatically calculates the hypothesis test for us. At a significance level of 5 percent the slope coefficient is statistically significant.

In more detail, we do a double sided t-test with a null hypothesis that the slope estimator is 0 i.e. there is no relation between purity and hydrocarbons. The alternative hypothesis is that this value is different than 0. In order to evaluate the this we use a students t-test with 2 degrees of freedom. The resulting

value of this t-test ends up being 3.3861194436304 and the correponding p-value of 0.00329112239545355. As we can see this p-value is close to zero and is less than than our significance figure of 0.05, so we can say that the slope estimator is indeed statistically significant.