Petroleum Spirit

May 7, 2020

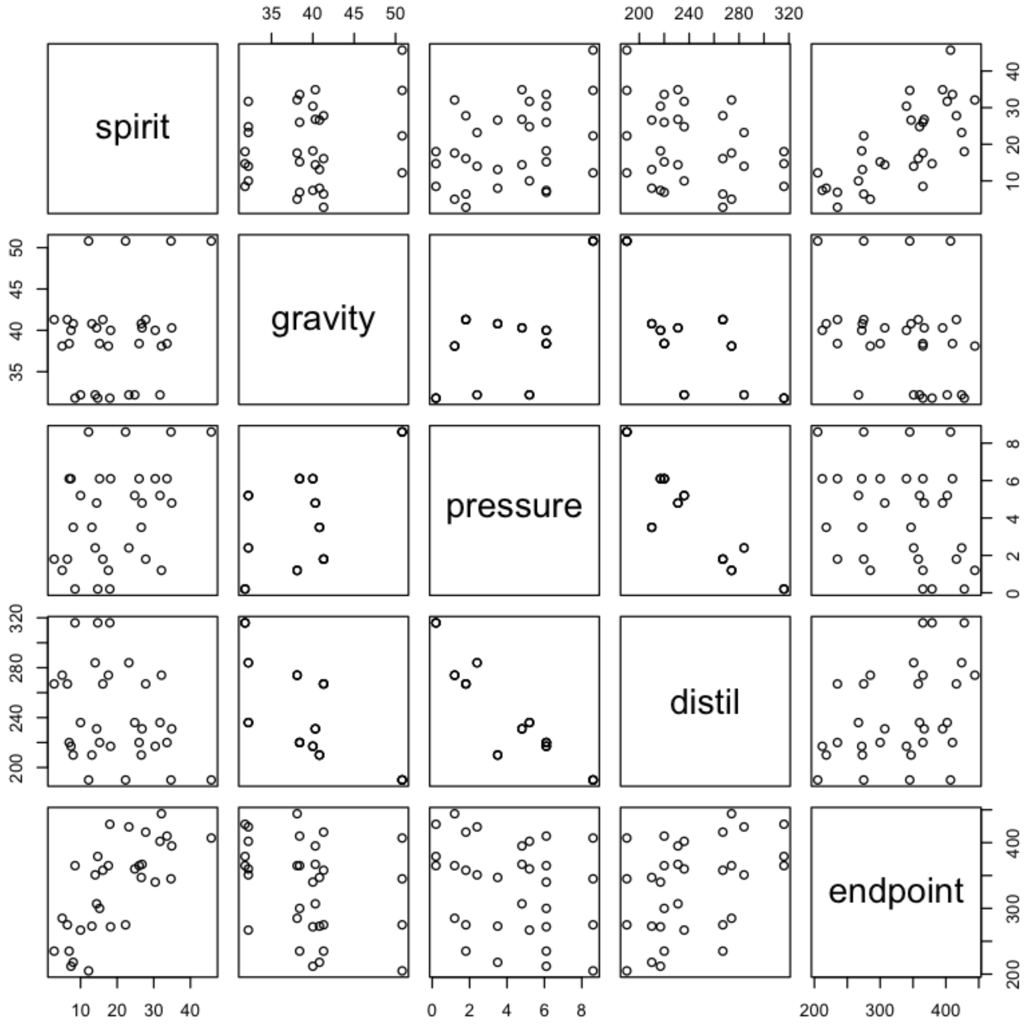
```
In [22]: df <- read.table('oil.txt', col.names = c('spirit','gravity','pressure','distil','endp

         df[1:5,]
```

spirit	gravity	pressure	distil	endpoint
6.9	38.4	6.1	220	235
14.4	40.3	4.8	231	307
7.4	40.0	6.1	217	212
8.5	31.8	0.2	316	365
8.0	40.8	3.5	210	218

0.1 Quick plot of the data

```
In [23]: pairs(df)
```



0.2 Fitting linear model

Starting by fitting the full model for all four explanatory variables

```
In [24]: dim(df)
```

1. 32 2. 5

```
In [38]: lm.full <- lm(spirit ~ gravity + pressure + distil + endpoint, data=df)
```

0.3 Linear model commentary

The number of observations is **32**. $n = 32$

We have four explanatory variables. Therefore $k = 4$ and $p = k + 1 = 5$

Therefore, SS_T, the total sum of squares has $n - 1$ degrees of freedom, 31

Recall $SS_T = SS_R + SS_{\text{Reg}}$

SS_R is the sum of squared residuals = $\mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$ which has $n - p$ degrees of freedom and follows a chi-squared distribution

Therefore the residual sum of squares has $32 - 5 = 27$ **degrees of freedom**.

SS_{Reg} is the regression sum of squares - a measure of the variation in the dependent variable captured by the regression model fit to the k explanatory variables, and has k degrees of freedom.

MS_R is the mean square residuals - **AND ALSO THE SAMPLE VARIANCE FOR Y / ϵ!** = and is calculated as $\dfrac{SS_R}{n-p}$

MS_{Reg} is the regression mean square, and is calculated as $\dfrac{SS_{Reg}}{k}$

0.3.1 Three hypothesis tests:

$$F_{k,n-p} = \frac{MS_R}{MS_{Reg}}$$

$$F_{k-m,n-p} = \frac{(SS_R(k) - SS_R(m))/(k-m)}{MS_{Reg}}$$

$$F_{1,n-p} = \frac{SS_R(k) - SS_R(k-1)}{MS_{Reg}}$$

```
In [39]: summary(lm.full)
```

```
Call:
lm(formula = spirit ~ gravity + pressure + distil + endpoint,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5804 -1.5223 -0.1098  1.4237  4.6214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.820774  10.123152  -0.674   0.5062
gravity      0.227246   0.099937   2.274   0.0311 *
pressure     0.553726   0.369752   1.498   0.1458
distil      -0.149536   0.029229  -5.116 2.23e-05 ***
```

```
endpoint     0.154650   0.006446  23.992  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.234 on 27 degrees of freedom
Multiple R-squared:  0.9622,Adjusted R-squared:  0.9566
F-statistic: 171.7 on 4 and 27 DF,  p-value: < 2.2e-16
```

0.4 Interpretation of these results

0.4.1 A few quick minor points:

- You can quickly suss out how many data points are in the sample by looking at the F-statistic
  - It's an $F_{k,n-p}$ statistic, so if you add $k+1$ to $n-p$ then you'll get n
  - So in the above, you've got a sample size of $n = 27 + 4 + 1 = 32$.
- Note that Residual standard error is equal to the square root of the residual mean square
  - We know $MS_{Reg} = s^2$, where s^2 is the sample estimate of the variance in the random error / dependent variable.
  - Therefore $\sqrt{MS_{Reg}} = s$ = Residual standard error

0.4.2 Looking coefficient by coefficient:

- gravity is significant at the 5% significance level, in the presence of all other variables
- pressure is **not** significant at the 5% significance level, in the presence of all other variables
- distil is highly significant
- endpoint is also highly significant

0.4.3 Looking at the overall F-statistic, which tests the null hypothesis that at least one of the estimated explanatory variable parameters is significantly non-zero

- We get $F_{k,n-p} = F_{4,27} = 171.7$

```
In [27]: pf(q = 171.7, df1 = 4, df2 = 27, lower.tail = FALSE)
```

```
8.82927612151701e-19
```

As we can see in both the linear model summary output, and the above calculation which doesn't windorise the p-value at 2e16, there is an extremely small probability of seeing a similar or more extreme F value and for the null to be true.

We therefore have sufficient evidence to reject the null hypothesis, and accept the null, that at least one of the explanatory variables is significant.

0.4.4 Looking at the ANOVA table, which looks a little different to the one in the lectures:

- Each explanatory variable has a regession sum of squares associated with it.
- SS_{Reg} is calculated by summing the individual coefficient sums of squares
- Each of these are associated with a single degree of freedom - you sum the 1's to produce k: the number of degrees of freedom associated with SS_{Reg}

0.4.5 ORDER MATTERS!!

- Each of these variables are added in sequence.
- Recall the delta F test of individual variables is to do:

$$\frac{SS_{Reg}(k) - SS_{Reg}(k-1)}{MS_R}$$

- Two routes to MS_R:

1) Use the square of the Residual standard error from the LM summary, 2.234**2 = 4.990756

2) Divide SS_R by the number of degrees of freedom, $n - p$, 134.80396 / 27 = 4.99273925925926

So we can calculate the $SS_{Reg}(k) - SS_{Reg}(k-1)$ bit by taking the full model **regression** sum of squares, and then subtracting the regression sum of squares of the model without the last variable:

```
In [28]: ss.reg.k <- sum(216.25577,309.85083,29.21432,2873.95231)
         ss.reg.k.sub.1 <- sum(216.25577,309.85083,29.21432)
```

```
In [29]: ss.reg.k - ss.reg.k.sub.1
```

```
2873.95231
```

0.5 This is of course just the SS_{Reg} associated with endpoint = 2873.95231

```
In [30]: F.1.27 <- (ss.reg.k - ss.reg.k.sub.1) / 4.992739
         F.1.27
```

```
575.626386638677
```

0.6 Taking the square of this, we expect to get the same t-statistic with 27 degrees of freedom as found in the LM summary

```
In [31]: sqrt(F.1.27)
```

```
23.9922151257169
```

0.6.1 endpoint 0.154650 0.006446 23.992 < 2e-16 ***

And that's exactly what we see here.

```
In [40]: anova(lm.full)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gravity	1	216.25577	216.255767	43.31405	4.643832e-07
pressure	1	309.85083	309.850828	62.06028	1.807364e-08
distil	1	29.21432	29.214317	5.85136	2.257822e-02
endpoint	1	2873.95231	2873.952314	575.62635	9.643907e-20
Residuals	27	134.80396	4.992739	NA	NA

0.7 Other comments on the ANOVA table

- The mean square regression statistics are calculated as the sum of squares divided by the number of degrees of freedom - which is 1 for the individual explanatory variables
- The F value associated with each variable will be the mean square regression statistics divided by the mean square residual.
- Those F values will therefore be $F_{1,27}$ statistics

```
In [33]: # Example of the gravity p-value associated with the F-statistic
         pf(43.31405, 1, 27, lower.tail = FALSE)
```

4.64383297501018e-07

0.8 Okay, so now we want to think about variable selection and drop variable from the model that aren't significant

0.8.1 Dropping pressure from the model

```
In [41]: lm.ex.pressure <- lm(spirit ~ gravity + distil + endpoint, data=df)

         summary(lm.ex.pressure)
```

```

Call:
lm(formula = spirit ~ gravity + distil + endpoint, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5303 -1.3606 -0.2681  1.3911  4.7658 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.032034   7.223341   0.558   0.5811    
gravity      0.221727   0.102061   2.173   0.0384 *  
distil      -0.186571   0.015922 -11.718 2.61e-12 ***
endpoint     0.156527   0.006462  24.224  < 2e-16 ***
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.283 on 28 degrees of freedom
Multiple R-squared:  0.959,Adjusted R-squared:  0.9546
F-statistic: 218.5 on 3 and 28 DF,  p-value: < 2.2e-16
```

```
In [42]: anova(lm.ex.pressure)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gravity	1	216.2558	216.255767	41.47339	5.648772e-07
distil	1	142.0838	142.083754	27.24873	1.519953e-05
endpoint	1	3059.7365	3059.736537	586.79425	2.522977e-20
Residuals	28	146.0011	5.214326	NA	NA

So the thing we totally expect - the R^2 reduces because that's $\frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_R}{SS_T}$, and the regression sum of squares will always increase or stay the same - never reduce after adding another variable

We're looking for the F-statistic to remain high, it's significance to remain as low, and for the R^2 / adjusted R^2 to not fall off too much, all whilst each individual parameter remains as significant in the presence of other variables

With this model, we have a highly significant F-statistic where we can confidently reject the null, plus three coefficients that are all significant at the 5% level of significance, and our R squared values look good.

This model would pass the initial goodness of fit tests, and we'd want to futher test out model adequecy by looking at plots of residuals to see how our assumptions of constant variance and mean zero error terms hold

We may also wish to go for a more parimmonious model by removing the gravity value - that's more of a subjective call - the benefits of one less parameter Vs the loss of fitting performance.

1 Last two things:

1. Compare two nested models
2. Prediction

1.1 1) Comparing two models via ANOVA

Let's compare the model with and without **endpoint**

```
In [37]: lm.ex.endpoint <- lm(spirit ~ gravity + pressure + distil, data=df)
In [45]: anova(lm.full, lm.ex.endpoint)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
27	134.804	NA	NA	NA	NA
28	3008.756	-1	-2873.952	575.6263	9.643907e-20

So the **Sum of Sq** here is the same as the $SS_R(k) - SS_R(k-1)$ bit, and the F as a result is the same as $F_{1,n-p} = \dfrac{SS_R(k) - SS_R(k-1)}{MS_{Reg}}$.

You could therefore work out that MS_R is equal to:

```
In [46]: 2873.952 / 575.6263
```

4.99273921292338

1.2 2) Prediction

By doing $E[\hat{y}]$ and $Var(\hat{y})$, we can quickly see that the standard error of the fitted value is: $s\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}$

R sorts this out for us via the predict command

```
In [53]: x <- data.frame(distil = 200, endpoint=400, gravity=10)

         confidence <- predict(lm.ex.pressure, x, se.fit = TRUE, interval = c('confidence'))

         prediction <- predict(lm.ex.pressure, x, se.fit = TRUE, interval = c('prediction'))
```

```
In [54]: confidence
```

$fit

fit	lwr	upr
31.54569	24.35379	38.73758

$se.fit 3.51096839732665

$df 28

$residual.scale 2.28348988377397

```
In [55]: prediction
```

$fit

fit	lwr	upr
31.54569	22.9665	40.12487

$se.fit 3.51096839732665

$df 28

$residual.scale 2.28348988377397

Both of these predictions use the same se.fit = $s\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}$

Confidence purely takes into account of standard errors in the parameter estimates (the reducible error)

Prediction takes into account both standard errors in the parameter estimates **AND** the error caused by the random error term in the model, ϵ_i.

```
In [ ]:
```