Midterm Comments

Statistics 135

Autumn 2005

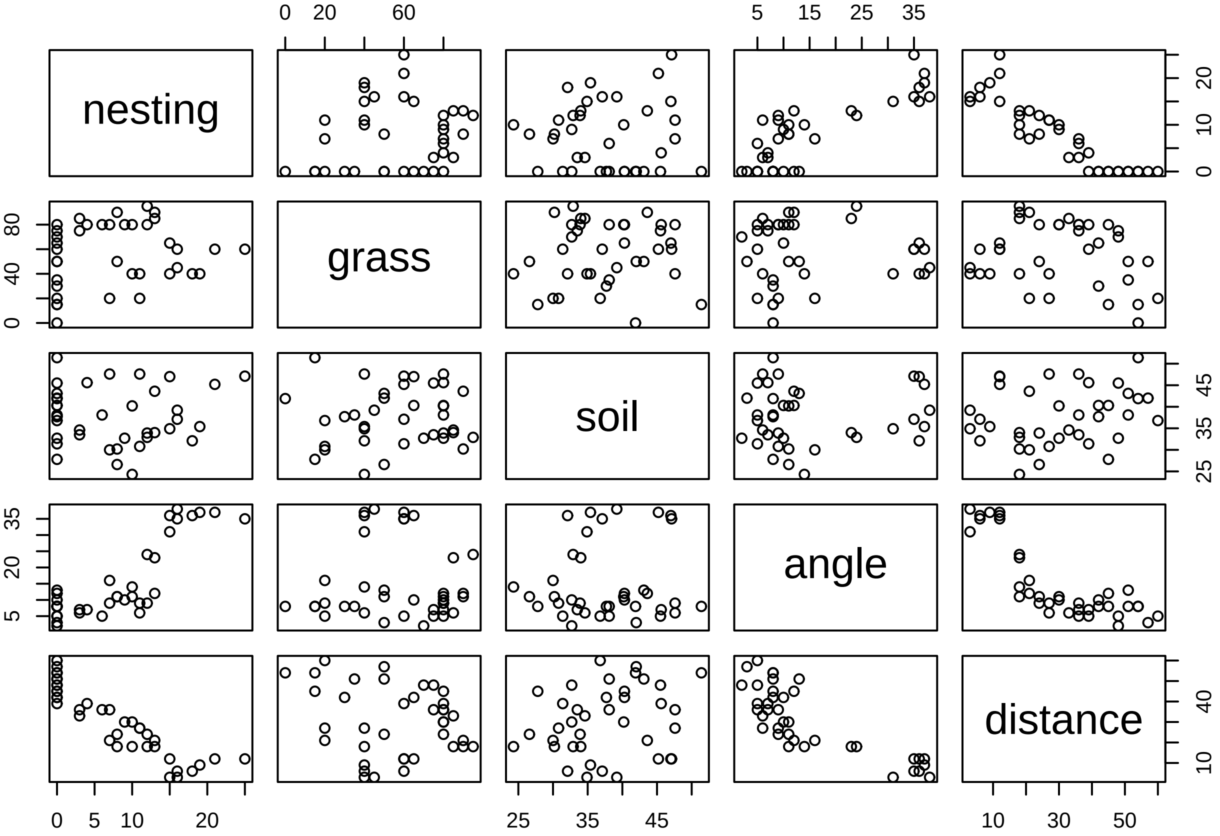
Midterm Comments

1. Puffin nesting frequency

$$\text{nesting}_i = \beta_0 + \beta_1 \text{grass}_i + \beta_2 \text{soil}_i + \beta_3 \text{angle}_i + \beta_4 \text{distance}_i + \epsilon_i$$

   (a) (5 points) The data from this study is available on the datasets page in the file `puffin.txt`. Read in the data, calculate the standard summary statistics (mean, standard deviation, and 5 figure summary) for each of the variables and create a scatter plot matrix of the data. Does it appear that any of the potential predictor variables are associated with nesting frequency?

```
> summary(puffin)
    nesting           grass            soil           angle
 Min.   : 0.000   Min.   : 0.00   Min.   :24.30   Min.   : 2.00
 1st Qu.: 0.000   1st Qu.:40.00   1st Qu.:32.75   1st Qu.: 7.25
 Median : 7.500   Median :60.00   Median :37.40   Median :10.00
 Mean   : 7.684   Mean   :56.45   Mean   :37.72   Mean   :15.00
 3rd Qu.:12.750   3rd Qu.:80.00   3rd Qu.:42.83   3rd Qu.:21.25
 Max.   :25.000   Max.   :95.00   Max.   :51.40   Max.   :38.00
    distance
 Min.   : 3.00
 1st Qu.:18.00
 Median :30.00
 Mean   :30.39
 3rd Qu.:44.25
 Max.   :60.00
> sapply(puffin, function(x) sqrt(var(x)))
  nesting     grass      soil     angle  distance
 7.192734 25.306443  6.653978 11.691993 16.581558
> sqrt(diag(var(puffin)))
  nesting     grass      soil     angle  distance
 7.192734 25.306443  6.653978 11.691993 16.581558
```



```
plot(puffin)                  # splom(~puffin) is fine as well
```

It appears that `angle` and `distance` are associated with `nesting` and possibly `soil` as well. Note that there are correlations between some of the predictors, particularly between `angle` and `distance`.

While not asked for, the correlation matrix for all the variables is

```
> cor(puffin)
             nesting       grass       soil       angle   distance
nesting   1.00000000  0.15848477 0.02167879  0.83558202 -0.9078590
grass     0.15848477  1.00000000 0.06935884 -0.01735530 -0.2052506
soil      0.02167879  0.06935884 1.00000000  0.06579734  0.2116592
angle     0.83558202 -0.01735530 0.06579734  1.00000000 -0.8146941
distance -0.90785903 -0.20525060 0.21165916 -0.81469413  1.0000000
```

(b) (5 points) Run the linear regression model for the above model and give the standard summaries (parameter estimates and standard errors, ANOVA table, etc). What evidence is there that some of the variables are useful in describing nesting frequency?

```
> puffin.lm <- lm(nesting ~ grass + soil + angle + distance, data=puffin)
> summary(puffin.lm)
Call:
lm(formula = nesting ~ grass + soil + angle + distance, data = puffin)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0166 -2.1088  0.2293  1.2505  6.9881

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.117840   3.185028   3.177  0.00323 **
grass       -0.007408   0.019459  -0.381  0.70586
soil         0.209211   0.077238   2.709  0.01062 *
angle        0.082389   0.077796   1.059  0.29727
distance    -0.366571   0.057473  -6.378 3.18e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.647 on 33 degrees of freedom
Multiple R-Squared: 0.8792,     Adjusted R-squared: 0.8645
F-statistic: 60.03 on 4 and 33 DF,  p-value: 1.113e-14
```

```
> anova(puffin.lm)
Analysis of Variance Table

Response: nesting
          Df  Sum Sq Mean Sq  F value    Pr(>F)
grass      1   48.08   48.08   6.8599   0.01321 *
soil       1    0.22    0.22   0.0313   0.86057
angle      1 1349.50 1349.50 192.5410 2.506e-15 ***
distance   1  285.12  285.12  40.6802 3.184e-07 ***
Residuals 33  231.29    7.01
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The clearest evidence is given by the F test from the `summary` output which examines the hypotheses

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_A : \text{at least one } \beta_i \neq 0, i = 1, \ldots, 4$$

```
F-statistic: 60.03 on 4 and 33 DF,  p-value: 1.113e-14
```

which is highly significant. At least one of the predictors appears to be useful in predicting nesting frequency. To determine the most likely ones are `distance` and `soil` based on the t tests from the `summary` output. After these two variables have been accounted for, the other two variables don't add much.

The F tests from the ANOVA table are not particularly useful in this case as they do not give tests that we are interested in this case.

Aside: Common Tests in Regression

For a regression model of the form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i$$

there are three common testing situations about the predictors

1. Testing all βs simultaneously:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_A : \text{at least one } \beta_i \neq 0, i = 1, \ldots, p$$

This is equivalent to comparing the models

$$\begin{array}{ll} \text{Reduced Model}(H_0): & y_i = \beta_0 + \epsilon_i \\ \text{Full Model}(H_A): & y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i \end{array}$$

2. Testing one β:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

This is equivalent to comparing the models (when $j = 1$)

$$\begin{array}{ll} \text{Reduced Model}(H_0): & y_i = \beta_0 + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i \\ \text{Full Model}(H_A): & y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i \end{array}$$

3. Testing a subset of the βs e.g.

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

The alternative hypothesis corresponds to the three situations

$$\begin{array}{ll} 1. & \beta_1 = 0, \beta_2 \neq 0, \beta_3, \ldots, \beta_p \text{ arbitrary} \\ 2. & \beta_1 \neq 0, \beta_2 = 0, \beta_3, \ldots, \beta_p \text{ arbitrary} \\ 3. & \beta_1 \neq 0, \beta_2 \neq 0, \beta_3, \ldots, \beta_p \text{ arbitrary} \end{array}$$

This is equivalent to comparing the models

$$\begin{array}{ll} \text{Reduced Model}(H_0): & y_i = \beta_0 + \beta_3 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i \\ \text{Full Model}(H_A): & y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i \end{array}$$

These situations can all be examined within the same framework. Let $SSE(m)$ be the error sum of squares for model m,

$$SSE(m) = \sum_{i=1}^{n} (y_i - \hat{y}_i(m))^2$$

where $\hat{y}_i(m)$ is the fitted value for observation i under model m. Now let $df(m)$ be the error degrees of freedom for model m,

$$df(m) = n - k(m)$$

where $k(m)$ is the number of predictors in model m.

The usual test statistic in all three cases is

$$F = \frac{(SSE(R) - SSE(F))/(df(R) - df(F))}{SSE(F)/df(F)}$$

which is compared to and F distribution with $df(R) - df(F)$ and $df(F)$ degrees of freedom. The numerator degrees of freedom $df(R) - df(F)$ is the number of parameters given in H_0.

This is exactly what the **S** `anova` function does when given two `lm` objects.

It is possible to show that in the case of testing a single β, the t given in the `summary` output is equivalent to the F test as

$$t^2 = F$$

and if $t \sim t_{df}$, then $F = t^2 \sim F_{1,df}$.

The output given by `anova(puffin.lm)` earlier doesn't test what we really want to do

```
> anova(puffin.lm)
Analysis of Variance Table
Response: nesting
          Df  Sum Sq Mean Sq  F value    Pr(>F)
grass      1   48.08   48.08   6.8599   0.01321 *
soil       1    0.22    0.22   0.0313   0.86057
angle      1 1349.50 1349.50 192.5410 2.506e-15 ***
distance   1  285.12  285.12  40.6802 3.184e-07 ***
Residuals 33  231.29    7.01
```

What **S** does in this case examines the following sequence of models

1. grass

```
          Df  Sum Sq Mean Sq  F value    Pr(>F)
grass      1   48.08   48.08   6.8599   0.01321 *
```

$$\begin{aligned}\text{Reduced Model}(H_0): &\quad y_i = \beta_0 + \epsilon_i \\ \text{Full Model}(H_A): &\quad y_i = \beta_0 + \beta_1 \texttt{grass}_i + \epsilon_i\end{aligned}$$

2. soil

```
          Df  Sum Sq Mean Sq  F value    Pr(>F)
soil       1    0.22    0.22   0.0313   0.86057
```

$$\begin{aligned}\text{Reduced Model}(H_0): &\quad y_i = \beta_0 + \beta_1 \texttt{grass}_i + \epsilon_i \\ \text{Full Model}(H_A): &\quad y_i = \beta_0 + \beta_1 \texttt{grass}_i + \beta_2 \texttt{soil}_i \epsilon_i\end{aligned}$$

3. angle

```
          Df  Sum Sq Mean Sq  F value    Pr(>F)
angle      1 1349.50 1349.50 192.5410 2.506e-15 ***
```

$$\begin{aligned}
&\text{Reduced Model}(H_0): && y_i = \beta_0 + \beta_1 \texttt{grass}_i + \beta_2 \texttt{soil}_i + \epsilon_i \\
&\text{Full Model}(H_A): && y_i = \beta_0 + \beta_1 \texttt{grass}_i + \beta_2 \texttt{soil}_i + \beta_3 \texttt{angle}_i + \epsilon_i
\end{aligned}$$

4. distance

```
          Df  Sum Sq Mean Sq  F value    Pr(>F)
distance   1  285.12  285.12  40.6802 3.184e-07 ***
```

$$\begin{aligned}
&\text{Reduced Model}(H_0): && y_i = \beta_0 + \beta_1 \texttt{grass}_i + \beta_2 \texttt{soil}_i + \beta_3 \texttt{angle}_i + \epsilon_i \\
&\text{Full Model}(H_A): && y_i = \beta_0 + \beta_1 \texttt{grass}_i + \beta_2 \texttt{soil}_i + \beta_3 \texttt{angle}_i \\
& && \qquad + \beta_4 \texttt{distance}_i + \epsilon_i
\end{aligned}$$

So the order that the variables are entered into the model affects the results given by `anova` **unless** `Var(X)` is a diagonal matrix, i.e.

$$\mathrm{Corr}(X_i, X_j) = 0 \quad \text{for all pairs } i \ \& \ j$$

A partial justification of this can be seen by looking at the following regressions

```
> puff.lm <- lm(nesting ~ distance, data=puffin)
> angle.lm <- lm(angle ~ distance, data=puffin)

> nestres <- resid(puff.lm)
> angleres <- resid(angle.lm)

> puffang.lm <- lm(nestres ~ angleres)
> summary(puffang.lm)

Coefficients:
              Estimate Std. Error   t value Pr(>|t|)
(Intercept) -2.045e-17  4.557e-01 -4.49e-17   1.0000
angleres     1.755e-01  6.811e-02     2.577   0.0142 *

> puffboth.lm <- lm(nesting ~ distance + angle, data=puffin)
> summary(puffboth.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.95582    2.44522   5.707 1.88e-06 ***
distance    -0.29297    0.04871  -6.015 7.39e-07 ***
angle        0.17554    0.06908   2.541   0.0156 *
```

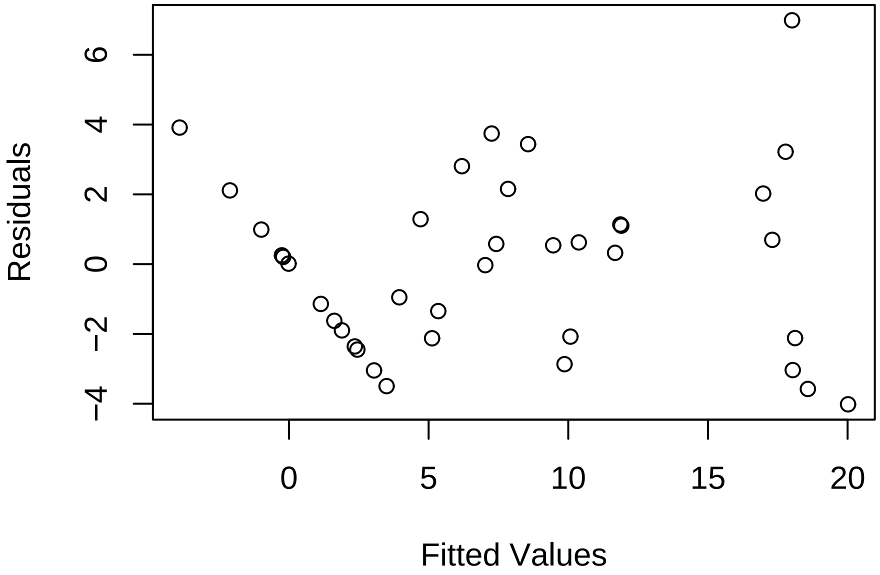
```
> puffa.lm <- lm(nesting ~ angle, data=puffin)
> summary(puffa.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02635    1.06591  -0.025     0.98
angle        0.51404    0.05633   9.126 6.75e-11 ***
```

So in multiple regression, the parameter estimate you get for a single predictor is the same as you get by regressing the residuals from a model with other predictors with the residuals you get from trying to predict the predictor of interest with the other predictors. (Sorry, I know this looks ugly).

So the bottom line is that what you get out of a multiple regression model for a variable, is what you get if you enter it last in a sequence of single variable regression models.

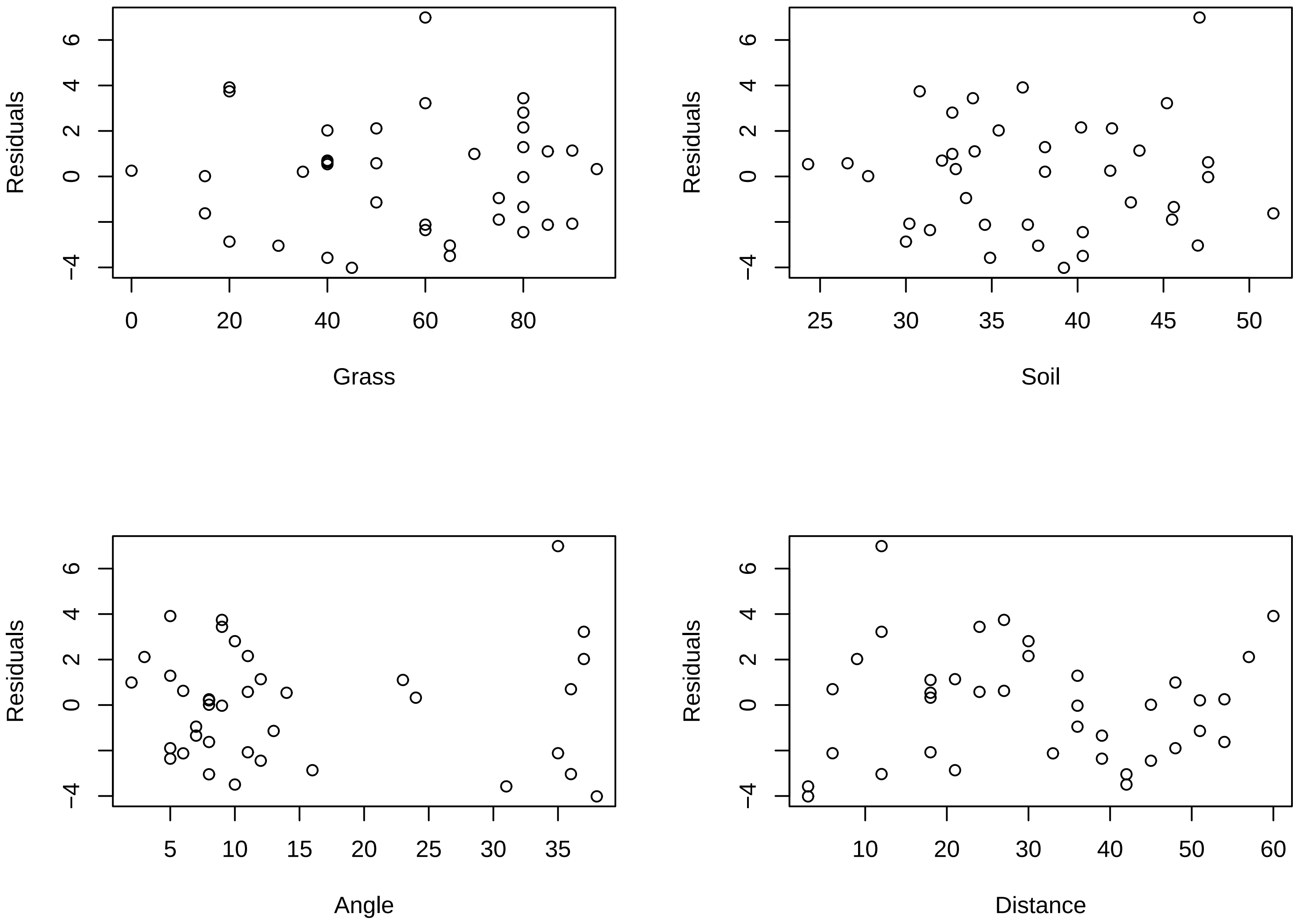
End Aside

(c) (5 points) Create the residual plot of the residuals against the fitted values. In addition, in a single figure, plot the residuals against each of the predictor variables. Do any of these figures suggest a problem with the regression model?



```
plot(fitted(puffin.lm), resid(puffin.lm), xlab="Fitted Values",
  ylab="Residuals")
```

Residuals vs Predictors



```
attach(puffin)
postscript("../Assignments/puffinres.eps", horiz=F, width=8, height=6.5)
par(mfrow=c(2,2), oma=c(0,0,4,0))
plot(grass, resid(puffin.lm), xlab="Grass", ylab="Residuals")
plot(soil, resid(puffin.lm), xlab="Soil", ylab="Residuals")
plot(angle, resid(puffin.lm), xlab="Angle", ylab="Residuals")
plot(distance, resid(puffin.lm), xlab="Distance", ylab="Residuals")
mtext(side=3, line=0, cex=1.5, outer=T, "Residuals vs Predictors")
```

The plots of the residuals versus the predictors don't show major problems. One outlier is suggested, though with 40 observations, not much to worry about, except for the fact that is it one the edge of the predictor space (high angle, low distance, high soil).

The plot of residuals versus the fits shows one problem, the decreasing line on the left of the plot. While it looks like curvature, that really isn't an accurate description of the problem. These are actually observations having the same value for nesting, in this case 0. It stands out in this case since it occurs with the minimum possible value of `nesting`.

(d) (5 points) Suppose that a new site was found where `grass` = 95, `soil` = 25, `angle` = 5, and `distance` = 60. Predict the number of nests for this site based on the original 38 sites. Any comments about this prediction?

```
> newdata <- data.frame(grass=95, soil=25,
                             angle=5, distance=60)
> predict(puffin.lm,newdata)
[1] -6.938013
```

Since the `nesting` can't be negative, this suggests a problem with the model. In fact we probably should have fit a Poisson regression model as the model fit allows for negative nesting values and the `nesting` values are counts.

(e) (5 points) Calculate the F test for comparing the model with all four predictors in the model with the model having only `soil` and `distance` in the model. What does this F test imply about the predictors `nesting`?

This test examines the hypotheses

$$H_0 : \beta_1 = \beta_3 = 0$$

$$H_A : \beta_1 \neq 0 \text{ or } \beta_3 \neq 0$$

```
> puffinred.lm <- lm(nesting ~ soil + distance, data=puffin)

> anova(puffinred.lm, puffin.lm)

Model 1: nesting ~ soil + distance
Model 2: nesting ~ grass + soil + angle + distance
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     35 244.869
2     33 231.293  2    13.576 0.9685 0.3902
```

In this case the p-value is large, suggesting that `grass` and `angle` do not add anything to the predictions after `soil` and `distance` have been accounted for.

(f) (5 points) Now fit the model

$$\text{nesting}_i = \beta_0 + \beta_1 \text{soil}_i + \beta_2 \text{distance}_i + \beta_3 \text{soil}_i \times \text{distance}_i + \epsilon_i$$

Is there any evidence of an interaction between distance and soil on the nesting frequency?

```
> puffinint.lm <- lm(nesting ~ soil * distance, data=puffin)

> summary(puffinint.lm)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.742393   5.679834   0.659  0.51440
soil           0.435530   0.148838   2.926  0.00608 **
distance      -0.157837   0.178889  -0.882  0.38380
soil:distance -0.006575   0.004535  -1.450  0.15625
```

```
> anova(puffinint.lm)
Analysis of Variance Table

Response: nesting
              Df  Sum Sq Mean Sq  F value Pr(>F)
soil           1    0.90    0.90   0.1326 0.7180
distance       1 1668.44 1668.44 245.9860 <2e-16 ***
soil:distance  1   14.26   14.26   2.1022 0.1563
Residuals     34  230.61    6.78
```

The test on the interaction (either t or F is fine here) clearly isn't significant

(g) (5 points) Now fit the model

$$\text{nesting}_i = \beta_0 + \beta_1\text{soil}_i + \beta_2\text{distance}_i + \beta_3\text{soil}_i^2 + \beta_4\text{distance}_i^2 + \epsilon_i$$

Is there any evidence of a nonlinearity in the relationship of distance or soil on the nesting frequency?

```
> puffinquad.lm <- lm(nesting ~ soil + distance + I(soil^2)
                      + I(distance^2), data=puffin)

> summary(puffinquad.lm)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.220574  14.035610   1.156 0.256114
soil           0.024878   0.732079   0.034 0.973095
distance      -0.495445   0.113372  -4.370 0.000116 ***
I(soil^2)      0.002714   0.009538   0.285 0.777772
I(distance^2)  0.001348   0.001826   0.738 0.465510
```

```
> anova(puffinquad.lm)
Response: nesting
              Df  Sum Sq Mean Sq  F value Pr(>F)
soil           1    0.90    0.90   0.1233 0.7277
distance       1 1668.44 1668.44 228.7155 <2e-16 ***
I(soil^2)      1    0.16    0.16   0.0223 0.8822
I(distance^2)  1    3.98    3.98   0.5452 0.4655
Residuals     33  240.73    7.29
```

Since both t tests on the quadratic terms are clearly not significant, the linearity of the relationship seems ok. However a better test of this (which agrees with the conclusion) is

```
> anova(puffinred.lm, puffinquad.lm)
Model 1: nesting ~ soil + distance Model 2: nesting ~ soil +
distance + I(soil^2) + I(distance^2)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     35 244.87
2     33 240.73  2      4.14 0.2837 0.7548
```

2. (30 points) The Ryan-Joiner test considers the following hypotheses

$$H_0 :\{x_1, x_2, \ldots, x_n\} \text{ is a random sample from a normal population}$$
$$H_A :\{x_1, x_2, \ldots, x_n\} \text{ is not a random sample from a normal population}$$

The test statistic is r, the correlation coefficient of the coordinate pairs,

$$\left(\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right), x_{(i)}\right), \quad i = 1, \ldots, n$$

of a normal Q-Q plot, where $\Phi^{-1}(\cdot)$ denotes the inverse CDF of a standard normal RV and $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ are the ordered data values. Under H_0, r should be close to 1.

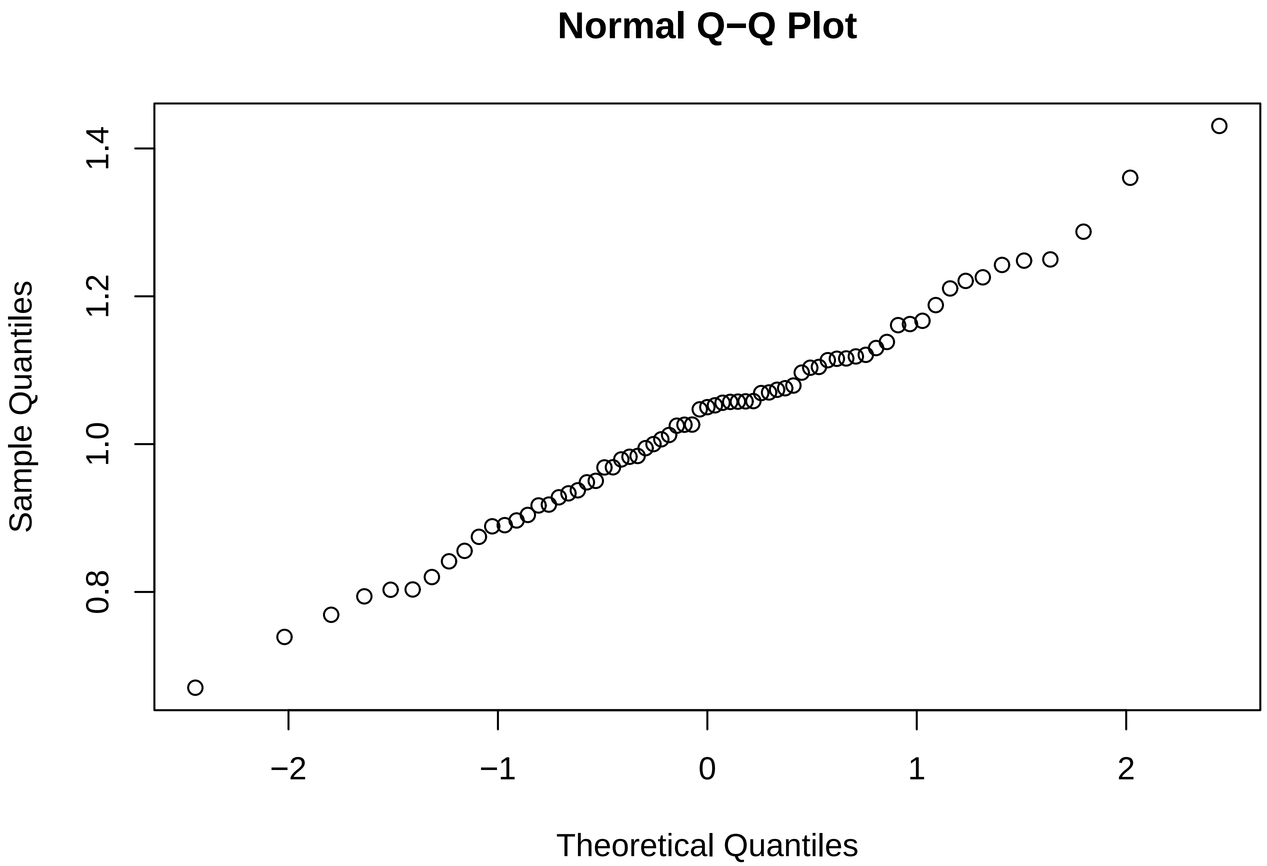
(a) (5 points) For the dataset `douglasfir.txt`, available on the datasets page for the course, calculate in **R**, the value of the statistic r described above.

```
rjstat <- function(x) {
   n <- length(x)
   cor(qnorm(((1:n)-0.375)/(n+0.25)),sort(x))
 }
>
> rjstat(douglasfir)
[1] 0.996571
```

(b) (5 points) Create the Q-Q plot (aka Normal Scores plot) for this dataset.

The plot can be created by

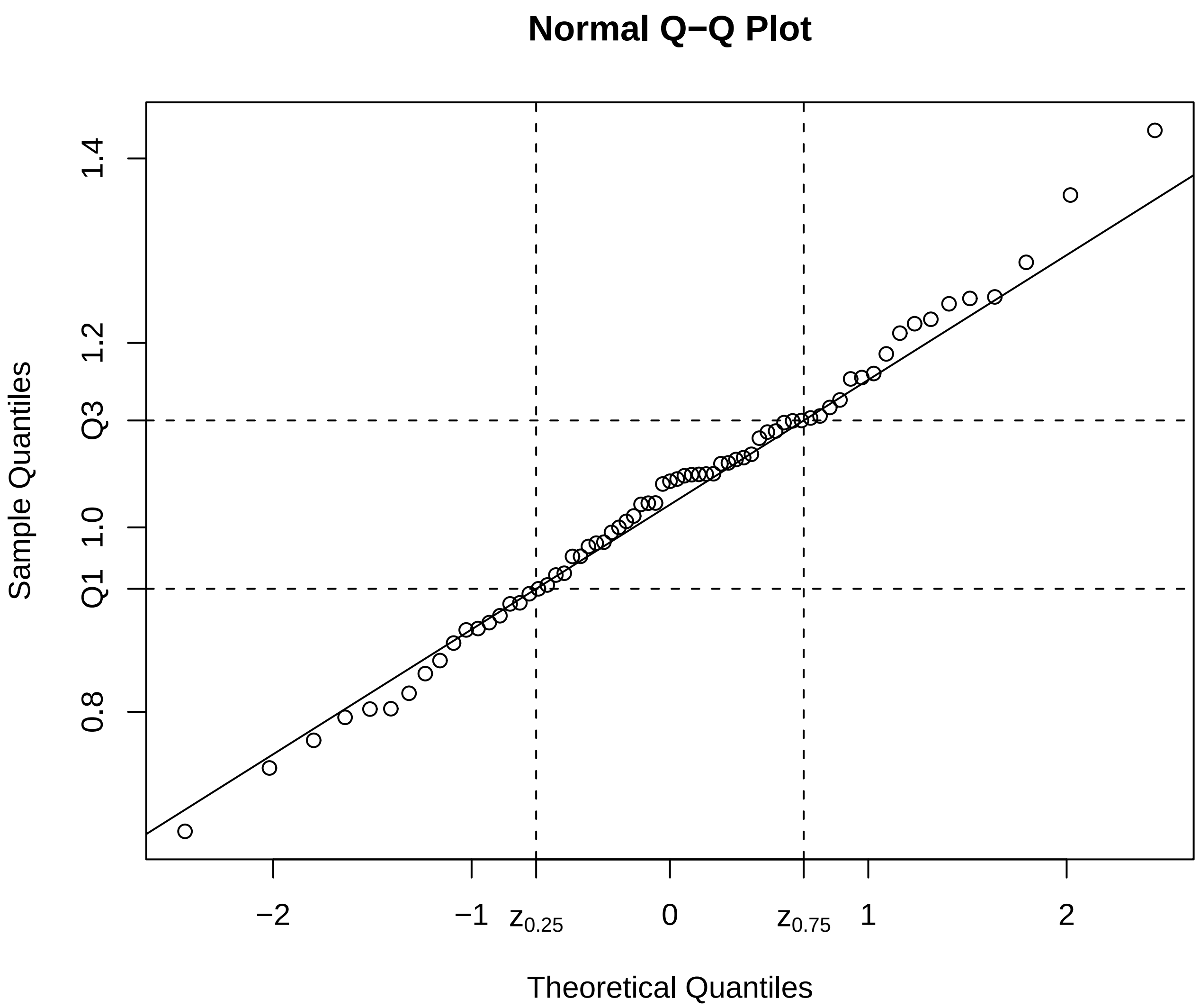
```
qqnorm(douglasfir)
```



(c) (5 points) The function `qqline` will add a straight line to Q-Q plots created by either `qqnorm` or `qqplot`. This line is a description of the main trend in the plot. By examining the code for `qqline`, what line is added to the `qqnorm` plot (i.e. when `datax = FALSE`).

```
> qqline
function (y, datax = FALSE, ...) {
    y <- quantile(y[!is.na(y)], c(0.25, 0.75))
    x <- qnorm(c(0.25, 0.75))
    if (datax) {
        slope <- diff(x)/diff(y)
        int <- x[1] - slope * y[1]
    }
    else {
        slope <- diff(y)/diff(x)
        int <- y[1] - slope * x[1]
    }
    abline(int, slope, ...)
}
```

Let $Q1$ and $Q3$ be the sample quartiles of the dataset and $z_{0.25}$ and $z_{0.75}$ be the population quartiles of the $N(0,1)$ distribution. Then `qqline` draws the line through the points $(z_{0.25}, Q1)$ and $(z_{0.75}, Q3)$. The slope of this line is an estimate of the standard deviation, though different than the sample standard deviation s.


Normal Q−Q Plot
Sample Quantiles
1.4
1.2
Q3
1.0
Q1
0.8
−2
−1
$z_{0.25}$
0
$z_{0.75}$
1
2
Theoretical Quantiles

(d) (10 points) Get p-value of test statistic by simulation

```
rjtest <- function(x, niter = 1000) {
  n <- length(x)
  r <- rjstat(douglasfir)
  h0mat <- matrix(rnorm(n*niter), ncol=n)
  rjh0 <- apply(h0mat, 1, rjstat)
  pval <- mean(rjh0 <= r)
  list(pval=pval, r=r, niter=niter)
}

> rjtest(douglasfir)
$pval [1] 0.934
$r [1] 0.996571
$niter [1] 1000

> rjtest(douglasfir,10000)
$pval [1] 0.9514
$r [1] 0.996571
$niter [1] 10000
```

(e) (5 points) Based on the above analysis, is there any evidence to suggest that the Douglas fir data is not normally distributed?

No. The Q-Q plot is very linear and the p-value from the Ryan-Joiner test is large.

One comment about calculating the p-value. One way of thinking of a p-value is

$$p\text{-value} = \text{P[a test stat as or more extreme than observed } |H_0]$$

In this case, only small values of r should be considered extreme. rs close to 1 are highly consistent with H_0, so a "one-sided" p-value should be calculated here, which is why

$$\hat{p} = \frac{1}{m}\sum_{j=1}^{m} I(r_j \leq r)$$

was used.