The Importance of Replication in Regression

The need for replication when determining an analyte’s concentration in a particular sample is obvious. Since any single analysis is subject to indeterminate error, the mean of several replicate analyses will always provide a better estimate of the analyte’s true concentration. In addition, using a t-test to compare the average experimental result for the analyte’s concentration to a known value requires knowing the standard deviation for the analysis, which requires replication.

Although not obvious, replication in regression also is important. Recall SPS5 in which you were asked to fit a straight-line model to some data and comment upon the model’s suitability. You may recall that most of you thought that the model was good because the data points were clustered near the regression line and because the correlation coefficient was high. We used the data to introduce the idea of examining the regression model in other ways, including looking at residual errors for evidence of no pattern, a normal distribution of values and the absence of leverage points. We also noted that statistical testing (an F-test of $MS_{\text{reg}}$ to $MS_{\text{res}}$) was limited to deciding whether there is a correlation between the dependent and independent variables, but not a decision about whether the model is appropriate. Replication makes this possible.

Recall the following equations for sum-of-squares:

- Total Sum of Squares: $SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SS_{\text{reg}} + SS_{\text{res}}$; (df = $n - 1$)
- Sum of Squares due to Regression: $SS_{\text{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$; (df = 1)
- Sum of Squares due to Residual Error: $SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$; (df = $n - 2$)

where $n$ is the number of data points used in determining the model. Now, suppose that we make replicate measurements for some or all $y_i$ and, where appropriate, calculate $\bar{y}_i$. This allows us to calculate two additional sum of squares:

- Sum of Squares due to Pure Error: $SS_{\text{pe}} = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$; (df = $n - k$)
- Sum of Squares due to Lack of Fit: $SS_{\text{lof}} = \sum_{i=1}^{n} (\bar{y}_i - \hat{y}_i)^2$; (df = $k - 2$)
where $k$ is the number of unique values for the independent variable. Note that the ‘2’ in the above equations for degrees of freedom reflects the two variables ($\beta_0$ and $\beta_1$) in the straight-line model. Note, also, that

$$SS_{res} = SS_{pe} + SS_{lof}$$

and that $SS_{pe}$ is a measure of indeterminate error. The $SS_{lof}$ includes contributions from both indeterminate error and determinate error. If $SS_{lof}$ is significantly larger than $SS_{pe}$, then we have evidence for a poor model.

Here is the raw data for the regression analysis

```r
> conc
[1] 0 1 1 2 2 3 3 4 4 5 5
> signal
[1] 0.00 0.98 0.90 2.10 2.20 3.16 3.22 3.68 3.72 4.15 4.27
```

and a plot of the data with a regression line.

```r
> plot(conc, signal)
> abline(lm(signal~conc))  # add the regression line
```

First we find the total sum of squares

```r
> avgy=mean(signal)  # calculate $\bar{y}$
> total=(signal-avgy)^2 # calculate $(y_i - \bar{y})^2$
> SStotal=sum(total); SStotal  # calculate SS_{total}
[1] 20.9902
```
Next, we find the sum of squares for regression and the sum of squares for residuals by first using the regression model to calculate values for $\hat{y}_i$

```r
> coef(lm(signal ~ conc))  
  (Intercept) conc  
  0.2651613  0.8487742
```

# find estimates for $\beta_0, \beta_1$

```r
> pred = 0.2651613 + 0.8487742 * conc  
```

# calculate values for $\hat{y}_i$

and then calculating the sum of squares

```r
> reg = (pred - avgy)^2  
> SSreg = sum(reg); SSreg  
[1] 20.30268
```

# calculate $(\hat{y}_i - \bar{y})^2$

```r
> res = (pred - signal)^2  
> SSres = sum(res); SSres  
[1] 0.6875213
```

# calculate $(y_i - \hat{y}_i)^2$

As expected, the SS\text{total} is equivalent to the sum of SS\text{reg} and SS\text{res}.

```r
> SStotal; SSreg + SSres  
[1] 20.9902
```

# SS\text{total} = SS\text{reg} + SS\text{res}

To determine if the model shows an appropriate correlation between the dependent and independent variables, we use the following hypotheses

$$H_0: \text{MS}_{\text{reg}} = \text{MS}_{\text{res}} \quad H_A: \text{MS}_{\text{reg}} \neq \text{MS}_{\text{res}}$$

The mean square terms are

```r
> MSreg = SSreg/1; MSreg  
[1] 20.30268
```

# MS\text{reg} with 1 dof

```r
> MSres = SSres/9; MSpe  
[1] 0.07639125
```

# MS\text{pe} with $11 - 2 = 9$ dof

The F-test is one-tailed; thus we calculate the value for $F$

```r
> MSreg/MSres  
[1] 265.7723
```

# $F_{\exp} = \text{MS}_{\text{reg}}/\text{MS}_{\text{res}}$

and then find the probability that $F_{\exp}$ can be this large
Because the probability is so small, we reject $H_0$ and accept $H_A$. We have strong statistical evidence that there is a correlation between the dependent and independent variables. We do have evidence, however, that the model is appropriate. To accomplish this we need to calculate $SS_{pe}$ and $SS_{lof}$.

To find the sum of squares for pure error and for lack of fit we first must create an object containing the average signal for each unique independent variable; thus

```r
> signal
[1] 0.00 0.98 0.90 2.10 2.20 3.16 3.22 3.68 3.72 4.15 4.27
> avgsig=c(0,0.94,0.94,2.15,2.15,3.19,3.19,3.70,3.70,4.21,4.21)

Thus

```r
> pe=(signal-avgsig)^2
> SSpe=sum(pe); SSpe
[1] 0.018

> lof=(avgsig-pred)^2
> SSlof=sum(lof); SSlof
[1] 0.6695213
```

As expected, the $SS_{res}$ is equivalent to the sum of $SS_{pe}$ and $SS_{lof}$

```r
> SSres; SSpe+SSlof # SS_{res} = SS_{pe} + SS_{lof}
[1] 0.6875213
[1] 0.6875213
```

To determine if the model is appropriate, we use the following hypotheses

$$H_0: MS_{lof} = MS_{pe} \quad H_A: MS_{lof} \neq MS_{pe}$$

The mean square terms are

```r
> MSlof=SSlof/4; MSlof
[1] 0.1673803
> MSpe=SSpe/5; MSpe
[1] 0.0036
```
The F-test is one-tailed; thus we calculate the value for F

\[
\text{Fratio} = \frac{\text{MSlof}}{\text{MSpe}} \quad \# \text{ } F_{\text{exp}} = \frac{\text{MSlof}}{\text{MSpe}}
\]

\[\text{Fratio} = 46.49453\]

and then find the probability that \( F_{\text{exp}} \) can be this large

\[\text{pf}(\text{MSlof}/\text{MSpe}, 4, 5, \text{lower.tail}=\text{FALSE})\]

\[\text{Fratio} = 0.0001682602\]

Because the probability is so small, we reject \( H_0 \) and accept \( H_A \). Now we have strong statistical evidence to support our earlier observation that a straight-line model is inappropriate for this data.

In summary, for \( k \) unique values of the independent variable and for \( n \) total measurements of the dependent variable:

\[
SS_{\text{total}} = n - 1 \text{ degrees of freedom}
\]

\[
SS_{\text{reg}} = 1 \text{ degree of freedom}
\]

\[
SS_{\text{res}} = n - 2 \text{ degrees of freedom}
\]

\[
SS_{\text{pe}} = n - k \text{ degrees of freedom}
\]

\[
SS_{\text{lof}} = k - 2 \text{ degrees of freedom}
\]

An F-test of \( \text{MS}_{\text{reg}} \) to \( \text{MS}_{\text{res}} \) evaluates whether there is a correlation between the dependent and independent variables and an F-test of \( \text{MS}_{\text{lof}} \) to \( \text{MS}_{\text{pe}} \) evaluates the validity of the model.