Tutorial 6: Linear Regression – Answers Sheet

Rob Nicholls – nicholls@mrc-lmb.cam.ac.uk

MRC LMB Statistics Course 2014

Task 1

1.

```
m1 = lm(eruptions~waiting,data=faithful)
summary(m1)
```

Intercept: -1.874016, coefficient of 'waiting': 0.075628, $R^2_{\text{adj}} = 0.8108$. The model does have significant utility. Both parameters are significantly different from zero. This information alone would lead to the conclusion that there is a linear relationship between the two variables (although model assumptions have not been tested).

2.

```
plot(eruptions waiting,data=faithful)
```

The data appear to form two clusters.

3.

```
lines(fitted(m1)~faithful$waiting)
```

The model seems reasonable, visually.

4.

```
plot(m1)
```

The model is not appropriate for describing the relationship between eruption duration and waiting time. This is particularly evident from the plot of residuals vs fitted values, which exhibits strong systematic relationships, implying that the model is inappropriate.

Task 2

1.

The dependent variable is Height, and the independent variable is Girth.

```
m1 = lm(Height∼Girth,data=trees)
summary(m1)
```

This model has significant utility. The parameter estimate is $\hat{\beta} = 1.0544$.

2.

The model cannot be used to predict Girth given Height. This is because the model predicts Height, not Girth. In particular, note that standard error estimates could not be achieved if trying to predict Girth given Height using a model designed to predict Height given Girth. An equivalent model to predict Height would have a different distribution of residuals. Indeed, the two optimisation problems solved in parameter estimation for the two models are different. Note that these would be equivalent if all observations perfectly on a line.

```
m2 = lm(Girth∼Height,data=trees))
summary(m2)
```

This model has significant utility. Only the coefficient of Height is significant. The estimate of the slope is $\hat{\beta} = 0.25575$. This is very different to the equivalent parameter estimate from model `m1`. This demonstrates how it is important to know exactly what question you want to ask before you try to draw conclusions – it is always important to use an appropriate model to address a particular question.

3.

```
m3 = lm(Girth∼0+Height,data=trees))
summary(m3)

plot(Height∼Girth,data=trees)
lines(fitted(m1)∼trees$Girth)
lines(trees$Height∼fitted(m2))
lines(trees$Height∼fitted(m3))
```

The lines are different because they represent different models, which correspond to the optimisation of different functions.

4.

```
m1 = lm(Volume∼Girth,data=trees)
m2 = lm(Volume∼Height,data=trees)
summary(m1)
summary(m2)
```

Looking at the model summaries, model `m1` is better than `m2`.

```
plot(m1)
```

The residuals appear to be correlated with fitted values, and there is one point of high influence.

```
m1 = lm(Volume[-31]~Girth[-31],data=trees)
summary(m1)
```

Parameter estimates do change. The change is reasonably substantial, but not so dramatic as to result in significantly different parameter values – the changes are approximately one σ in magnitude.

Task 3

1.

```
cor(anscombe$x1,anscombe$y1)
cor(anscombe$x2,anscombe$y2)
cor(anscombe$x3,anscombe$y3)
cor(anscombe$x4,anscombe$y4)
```

All correlations are essentially the same.

2.

```
summary(lm(anscombe$y1~anscombe$x1))
summary(lm(anscombe$y2~anscombe$x2))
summary(lm(anscombe$y3~anscombe$x3))
summary(lm(anscombe$y4~anscombe$x4))
```

The summaries suggest that all models are essentially the same.

3.

```
plot(anscombe$y1~anscombe$x1)
plot(anscombe$y2~anscombe$x2)
plot(anscombe$y3~anscombe$x3)
plot(anscombe$y4~anscombe$x4)
```

Despite having the same correlations, and the same parameters and statistics when creating linear models, the relationships between x and y for each of these datasets are very different. This shows how it is vitally important to visually inspect the data, and ensure that model assumptions are satisfied, rather than naively drawing conclusions based on limited statistics.

Task 4

1.

```
plot(speed~dist,data=cars)
```

The relationship seems reasonably linear, but also seems to plateau.

2.

```
m1 = lm(speed~dist,data=cars)
summary(m1)
```

```
plot(m1)
```

The model does have significant utility. There is a dependency between residuals and fitted values.

3.

```
m2 = lm(speed~sqrt(dist),data=cars)
summary(m2)
plot(m2)
```

The relationship seems linear, and this model fits better than the previous model (according to R^2). Also, the dependency between the residuals and fitted values is reduced.

4.

```
m3 = lm(log(speed)~log(dist),data=cars)
summary(m3)
plot(m3)
```

This model fits better than the previous model (according to R^2). Consequently, it may be concluded that this model is better than the previous model.

5.

The model of speed against the square root of dist includes an intercept parameter that is not significantly different from zero. Consequently, this term may be removed from the model.

6.

```
plot(log(speed)~log(dist),data=cars)
lines(fitted(m3)~log(dist),data=cars)
```

7.

```
plot(speed~dist,data=cars)
x = order(cars$dist)
lines(exp(fitted(m3))[x]~dist[x],data=cars)
```

8.

```
lines(fitted(m2)[x]~dist[x],data=cars,col="red")
```

The curves are similar, but they are different. Note that the fitted values from the log-log model are systematically lower, as the model forces the curve to have an intercept of zero (pass through the origin).

Task 5

1.

```
plot(time∼conc,data=Indometh)
```

The relationship between time and conc is negative and non-linear. Specifically, time is inversely related to conc.

2.

```
plot(log(time)∼log(conc),data=Indometh)
```

3.

```
m1 = lm(log(time)∼log(conc),data=Indometh)
summary(m1)
plot(m1)
```

Intercept: -0.4203, coefficient of 'log(conc)': -0.9066. Both parameters are significantly different from zero. There are no concerning outliers with particularly large influence.

4.

```
plot(log(time)∼log(conc),data=Indometh)
lines(fitted(m1)∼log(conc),data=Indometh)
```

5.

```
plot(time∼conc,data=Indometh)
x = order(Indometh$conc)
lines(exp(fitted(m1))[x]∼conc[x],data=Indometh)
```

Task 6

1.

```
m1 = lm(Volume∼Height,data=trees)
m2 = lm(Volume∼Girth,data=trees)
summary(m1)
summary(m2)
```

The model summaries indicate that the model that regresses Volume on Girth is better, on the basis of a much higher Multiple R-squared value (0.9353 versus 0.3579).

2.

```
m3 = lm(Volume∼Height+Girth,data=trees)
summary(m3)
```

```
plot(m3)
```

This model is better than both of the previous models, on the basis that it has a higher Multiple R-squared value (0.948). All model terms are significant. The model seems reasonable – the diagnostic plots do not indicate anything particularly concerning about this model.

3.

```
m4 = lm(Volume∼Height*Girth,data=trees)
summary(m4)
plot(m4)
```

This model is better than the previous models, on the basis that it has a higher Multiple R-squared value (0.9756). Again, all model terms are significant, and the model seems reasonable – the diagnostic plots do not indicate anything particularly concerning about this model.

4.

```
m5 = lm(log(Volume)∼log(Height)+log(Girth),data=trees)
summary(m5)
plot(m5)
```

This model is better than the previous models, on the basis that it has a higher Multiple R-squared value (0.9777). Note also that this model has fewer parameters than the previous model. Again, all model terms are significant, and the model seems reasonable – the diagnostic plots do not indicate anything particularly concerning about this model. We can conclude that the relationship between Volume, Height and Girth is not linear. Rather, it is a power relationship, as it better described by a log-log model.

5.

```
m6 = lm(log(Volume)∼log(Height)*log(Girth),data=trees)
summary(m6)
plot(m6)
```

This model has a higher Multiple R-squared value than previous models (0.9778). However, this is at the expense of more parameters – more terms are present in the model. In fact, none of the parameter estimates are significantly different from zero. This means that not all terms are required in order to model log(Volume). If log(Height) and log(Girth) are present, then including the interaction term log(Height):log(Girth) does not result in a better model.

6.

The model that regresses log(Volume) on log(Height) and log(Girth) is the most appropriate.

```
plot(Volume∼Girth,data=trees)
```

```
x = order(trees$Girth)
lines(exp(fitted(m5))[x]~Girth[x],data=trees)
```

Task 7

1.

```
plot(conc~rate,data=Indometh)
```

The relationship between conc and rate is positive and non-linear. Visually, it appears that there may be a power or exponential relationship between conc and rate.

2.

```
plot(log(conc)~rate,data=Indometh)
m1 = lm(log(conc)~rate,data=Indometh)
summary(m1)
```

A logarithmic transformation of the dependent variable linearises the model and stabilises the variance, resulting in a linear model in which both terms are significant.

3.

```
m2 = lm(log(conc)~rate+state,data=Indometh)
summary(m2)
```

State is a factor variable with two levels. Including this term improves the model – the Multiple R-squared value increases, and all terms are significant.

4.

```
m3 = lm(log(conc)~rate*state,data=Indometh)
summary(m3)
```

Including the interaction term results in the state term not being significant. Consequently, not all three terms are required in the model.

5.

```
m4 = lm(log(conc)~rate*state-state,data=Indometh)
summary(m4)
```

All three terms in this model are required, and the Multiple R-squared value is higher than for the previous models m1 and m2.

```
plot(conc~rate,data=Puromycin)
x = Puromycin$state==levels(Puromycin$state)[1]
y = order(Puromycin$rate[x])
```

```
z = order(Puromycin$rate[!x])
lines(exp(fitted(m4))[x][y]~rate[x][y],data=Puromycin)
lines(exp(fitted(m4))[!x][z]~rate[!x][z],data=Puromycin,col="red")
```

Task 8

1.

```
m1 = lm(Fertility~.,data=swiss)
summary(m1)
```

Not all terms are significant.

2.

```
step(m1)
```

The Examination term was removed. The AIC corresponding to the final model is 189.86.

```
m2 = lm(Fertility~.-Examination,data=swiss)
summary(m2)
```

All terms in the final model are significant.

Task 10

The simple linear regression equation for an observation i is:

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \varepsilon_i \tag{1}$$

Summing over all observations, we get:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \hat{\alpha} + \sum_{i=1}^{n} \hat{\beta} x_i + \sum_{i=1}^{n} \varepsilon_i \\
&= \hat{\alpha} \sum_{i=1}^{n} 1 + \hat{\beta} \sum_{i=1}^{n} x_i \\
&= \hat{\alpha} n + \hat{\beta} \sum_{i=1}^{n} x_i \\
\frac{1}{n} \sum_{i=1}^{n} y_i &= \hat{\alpha} + \hat{\beta} \frac{1}{n} \sum_{i=1}^{n} x_i \\
\bar{y} &= \hat{\alpha} + \hat{\beta} \bar{x} \\
\hat{\alpha} &= \bar{y} - \hat{\beta} \bar{x}
\end{aligned} \tag{2}$$

Rearranging Equation (1) we get:

$$\varepsilon_i = y_i - \hat{\alpha} - \hat{\beta} x_i \tag{3}$$

Combining Equations (2) and (3) we get:

$$\begin{aligned}\varepsilon_i &= y_i - (\bar{y} - \hat{\beta}\bar{x}) - \hat{\beta}x_i \\ &= (y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\end{aligned}$$

The parameters $\hat{\alpha}$ and $\hat{\beta}$ are chosen so as to minimise the sum of squares of the residuals:

$$\begin{aligned}S &= \sum_{i=1}^{n} \varepsilon_i^2 \to \min \\ &= \sum_{i=1}^{n} \Big((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\Big)^2 \to \min\end{aligned}$$

This equation is minimised when the derivative of S with respect to $\hat{\beta}$ is equal to zero:

$$\begin{aligned}\frac{\partial S}{\partial \hat{\beta}} = -2\sum_{i=1}^{n}(x_i - \bar{x})\Big((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\Big) &= 0 \\ \sum_{i=1}^{n}(x_i - \bar{x})\Big((y_i - \bar{y}) - \hat{\beta}(x_i - \bar{x})\Big) &= 0 \\ \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= 0\end{aligned}$$

Rearranging this for $\hat{\beta}$:

$$\begin{aligned}\hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}\end{aligned}$$

And $\hat{\alpha}$ can be achieved by plugging this back into Equation (2):

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$