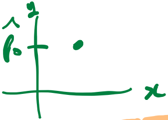
STA302 weeks 9–10

Mark Ebden 2018. Chapter 5

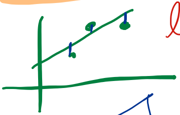
With grateful acknowledgment to Alison Gibbs



1 parameter : β_0

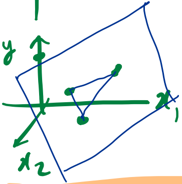
RSS=0

2 parameters : β_0, β_1



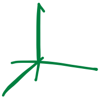
line.

3 parameters : $\beta_0, \beta_1, \sigma^2$



RSS=0
plane

or $\beta_0, \beta_1, \underline{\beta_2}$

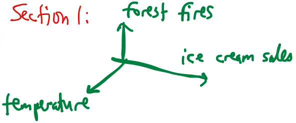


hyperplane

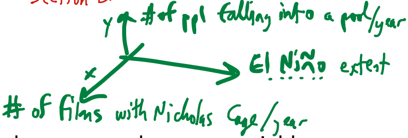
$p+2$ parameters

Multiple regression

Section 1:



Section 2:



Multiple regression is used when we have more than one explanatory variable. Multiple x's can arise naturally. In addition, sometimes we want to:

- Control for some x's to consider the effect on y of other x's over and above the control variables
- Fit a polynomial
- (Compare the regression line for two or more groups)

In multiple linear regression (MLR), generally we let p represent the number of explanatory variables in the model, i.e.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i$$

for $i \in \{1, \ldots, n\}$. How many parameters do we need to estimate? $p+2$

(using Gauss-Markov conditions)

And therefore, how many observations do we need at a minimum? $p+2$

parameter: σ^2 β_1 . . .

estimator: S^2 B_1

estimate: s^2 b_1

Matrix version of MLR

Our main equation is unchanged: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$

However, the **design matrix X** and β are bigger:

SLR

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & & X_{2p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

SLR

A design matrix gives the explanatory variables (often without the column of 1's). Each row is an observation and each column corresponds to a different kind of variable.

Gauss-Markov assumptions for MLR

The key equations are unchanged:

$$\mathrm{E}(\mathbf{e}) = \mathbf{0} \quad \text{and} \quad \mathrm{var}(\mathbf{e}) = \sigma^2 \mathbf{I}$$

For our inference methods (CIs etc), we need $\mathbf{e}$ to have a multivariate normal distribution as before.

$t_{n-2} \rightarrow t_{n-(p+1)}$

The expression for residuals is still $\widehat{\mathbf{e}} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, where now we have

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

Regression diagnostics

Leverage pts:

Cutoff $\frac{2(p+1)}{n}$

Estimating σ^2 in MLR

Recall that

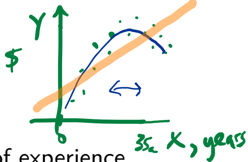
$$S^2 = \text{MSE} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\text{d.f. of error}} = \frac{\widehat{\mathbf{e}}'\widehat{\mathbf{e}}}{\text{d.f. of error}}$$

The number of degrees of freedom was $n-2$ in SLR, and is $n-p-1$ in MLR. To see this, recall that $\text{RSS} = \widehat{\mathbf{e}}'\widehat{\mathbf{e}} = \mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}$. Using our five key properties of idempotent matrices again, $\text{rank}(\mathbf{I}-\mathbf{H}) = \text{rank}(\mathbf{I}) - \text{rank}(\mathbf{H}) = n-(p+1)$ assuming that the columns of X are linearly independent.

To show that S^2 is unbiased in MLR, similar to before we can show $\mathrm{E}(\text{RSS}) = (n-p-1)\sigma^2$. The proof is akin to the SLR proof except that:

$$\begin{aligned}
\text{trace}(\mathbf{I}-\mathbf{H}) &= \text{trace}(\mathbf{I}) - \text{trace}(\mathbf{H}) \\
&= n - \text{trace}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \\
&= n - \text{trace}\left[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right] \\
&= n - \text{trace}(\mathbf{I}_{p+1}) \quad \text{2x2 in SLR} \\
&= n-(p+1)
\end{aligned}$$

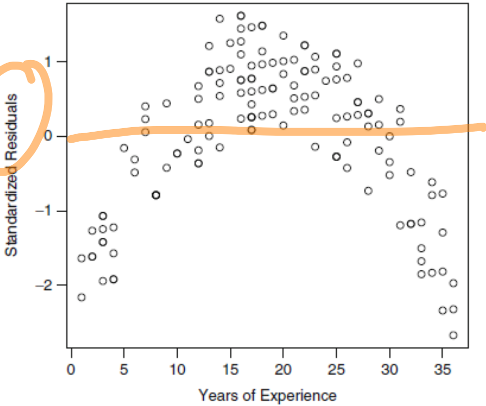
Example of MLR: Fitting a polynomial



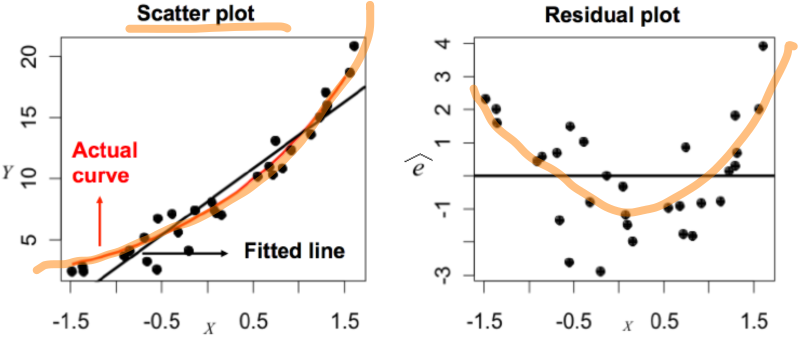
A professional-salary database contains 143 ordered pairs:

(years of experience, salary)

Generally, but not monotonically, salary increases with years of experience. Using SLR, our model is $Y_i = \beta_0 + \beta_1 x_i + e_i$. After fitting a straight line, this is the plot of standardized residuals:



Example of a nonlinear relationship *(Weeks 4–5, Slide 43)*



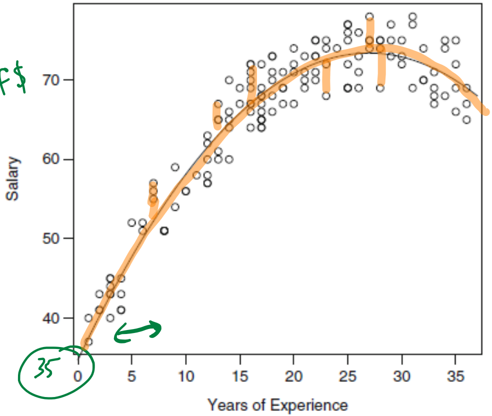
Remedial measure: If the regression function isn't linear,

- In some cases, a variable transformation can make the data "more linear"
- Otherwise, a different (e.g. nonlinear) model might be better

Back to our salary database

A simple nonlinear model is MLR in which we fit a parabola, i.e. incorporate x and x^2. The model is $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$ and the plot is:

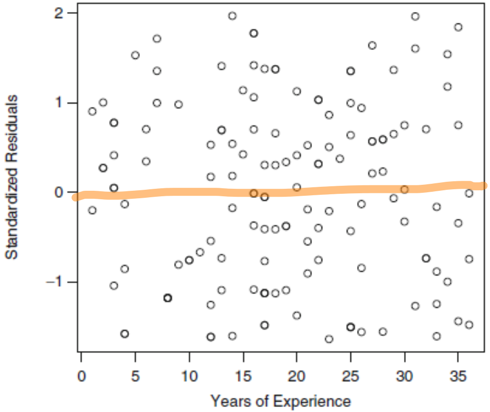
1000's of $



Here, $\hat{\beta}_0 \approx 35$, $\hat{\beta}_1 \approx 2.87$, and $\hat{\beta}_2 \approx -0.053$, each with $p < 2 \times 10^{-16}$.

MLR example: fitting a polynomial

The residuals no longer have a pattern:



R code for MLR

type: ?lm

```
X <- read.csv("profsalary.txt",sep="\t")
mod1 <- lm(Salary ~ Experience + I(Experience^2), data=X)
summary(mod1)
```

$y \sim x$

A 4 ^

Typing `I(.)` is a way to express formulae within a call to `lm`.

The + sign indicates that more than one explanatory variable is being used. To have four variables, use e.g. `y ~ x1 + x2 + x3 + x4`

$\cancel{y = \beta_0 + x_1 + x_2 + x_3 + x_4}$

no

R output for MLR

```
##
## Call:
## lm(formula = Salary ~ Experience + I(Experience^2), data = X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5786 -2.3573  0.0957  2.0171  5.5176
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.720498   0.828724   41.90   <2e-16 ***
## Experience       2.872275   0.095697   30.01   <2e-16 ***
## I(Experience^2) -0.053316   0.002477  -21.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.817 on 140 degrees of freedom
## Multiple R-squared:  0.9247, Adjusted R-squared:  0.9236
## F-statistic: 859.3 on 2 and 140 DF,  p-value: < 2.2e-16
```

new

Using the R model

Interpolate at 5 years of experience: $1 \quad x \quad x^2$

$\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2$

```
e <- 5; mod1$coefficients%*%c(1,e,e^2)
```

matrix mult

```
##          [,1]
## [1,] 47.74897
```

Alternatively, use the predict command:

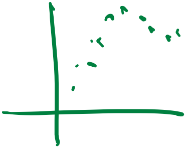
```
predict(mod1,data.frame(Experience=5))
```

```
##        1
## 47.74897
```

The data frame passed to predict names and initializes all of the information used towards making the predictor variables. Another example would be:

```
predict(lm(y~x1+x2),data.frame(x1=5,x2=3))
```

Interpreting MLR coefficients



How should we interpret β_j, or similarly their estimates b_j — i.e. what's the meaning of the coefficients of MLR predictor variables?

In general, β_j is the change in the mean value of Y associated with a one-unit change in the predictor variable x_j, with *all other variables held constant*.

For our salary database example, this is impossible. The closest interpretations we can make are of this sort:

- If Experience increases from 5 years to 6 years, the estimated change in mean Salary is $2.87 - 0.053(36 - 25) \approx 2.3$
- If Experience increases from 35 years to 36 years, the estimated change in mean Salary is $2.87 - 0.053(36^2 - 35^2) \approx -0.9$

Do we need a polynomial fit?

We can quantify whether the quadratic term is 0 or not using familiar hypothesis testing:

$$H_0 : \beta_2 = 0 \quad \text{vs} \quad H_a : \beta_2 \neq 0$$

Exercise: Try this on the salary database. What do you find?

Do we need the jth predictor?

β_j

In general, a test of $H_0 : \beta_j = 0$ gives an indication of whether or not the jth predictor variable statistically significantly contributes to the estimation/prediction of Y *over and above* the other predictor variables.

That is, the test assumes that the other variables are in the model.

Recap of Regression ANOVA *(Week 3)*

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n} b_1^2(x_i-\bar{x})^2}_{\text{SSReg}} + \underbrace{\sum_{i=1}^{n}\hat{e}_i^2}_{\text{RSS}}$$

Source	SS	d.f.	MS = SS/df
Regression line	$b_1^2 S_{xx} = \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$b_1^2 S_{xx}$
Error	$\sum_{i=1}^{n}\hat{e}_i^2$	$n-2$	S^2
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	–

The coefficient of determination is $R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}, \quad 0 \le R^2 \le 1.$

This week we showed that the ANOVA identity can be rewritten as:

$$\underbrace{\mathbf{Y}'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\text{SST}} = \underbrace{\mathbf{Y}'\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\text{SSReg}} + \underbrace{\mathbf{Y}'\left(\mathbf{I}-\mathbf{H}\right)\mathbf{Y}}_{\text{RSS}}$$

Introducing Multiple-Regression ANOVA

SLR → MLR

In multiple regression, the ANOVA identity is the same as before, albeit with a different $\mathbf{H} = \underset{n\times(p+1)}{\mathbf{X}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$:

$$\mathrm{SST} = \mathrm{SSReg} + \mathrm{RSS}$$

$$\underbrace{\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\mathrm{SST}} = \underbrace{\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}_{\mathrm{SSReg}} + \underbrace{\mathbf{Y}'\left(\mathbf{I} - \mathbf{H}\right)\mathbf{Y}}_{\mathrm{RSS}}$$

The MLR ANOVA table is similar to before, but the degrees of freedom change:

Source	SS	d.f.	MS = SS/df
Regression	SSReg	p	SSReg/p
Error	RSS	$n-p-1$	S^2
Total	SST	$n-1$	–

The F-test in an MLR ANOVA table

The test hypotheses are:

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- H_a : At least one of the β_j's isn't 0

The test statistic is:

$$F_{\text{obs}} = \frac{\text{MSReg}}{\text{MSE}}$$

1, n-2 SLR

If H_0 is true, F_{obs} is an observation from an F distribution with $(p,\, n-p-1)$ MLR degrees of freedom.

- Numerator d.f.: the # of β's being tested
- Denominator d.f.: the d.f. for the error

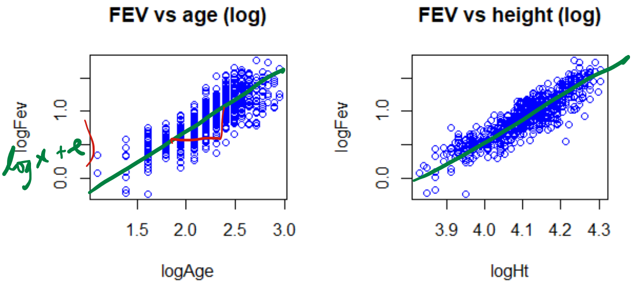
So in MLR ANOVA, we use the F-test to check for linear association between Y and *any* of the p predictors. If the F-test is significant, then we might ask, for *which* predictor(s) is there evidence of a linear association with Y? Some pitfalls in answering this question are investigated in Chapter 7.

Example of an F-test: the `fev` database

```r
a2 = read.table("DataPPC.txt",sep=" ",header=T) # Load the data set
logFev <- log(a2$fev); logAge <- log(a2$age); logHt <- log(a2$ht)
par(mfrow=c(1,2))
plot(logAge,logFev,type="p",col="blue",pch=21, main="FEV vs age (log)")
plot(logHt,logFev,type="p",col="blue",pch=21, main="FEV vs ht (log)")
mod1 = lm(logFev~logAge+logHt)
```

transform

MLR



$e^{(\log Y)} = e^{(\beta_0 + \beta_1 \log x + e)}$

$Y \propto x^{\beta_1}$

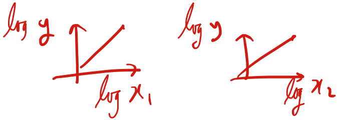
SLR in the `fev` database

```
##
## Call:
## lm(formula = logFev ~ logAge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60857 -0.13532  0.00227  0.14329  0.56348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.98772    0.05756  -17.16   <2e-16 ***
## logAge       0.84615    0.02535   33.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2026 on 652 degrees of freedom
## Multiple R-squared:  0.6309, Adjusted R-squared:  0.6303
## F-statistic:  1114 on 1 and 652 DF,  p-value: < 2.2e-16
```

SLR in the fev database

```
##
## Call:
## lm(formula = logFev ~ logHt)
##
                  y
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69369 -0.09122  0.01145  0.09832  0.44965
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.92110    0.25577  -46.61   <2e-16 ***
## logHt         3.12418    0.06223   50.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1512 on 652 degrees of freedom
## Multiple R-squared:  0.7945, Adjusted R-squared:  0.7941
## F-statistic:  2520 on 1 and 652 DF,  p-value: < 2.2e-16
```

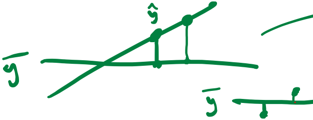
MLR in the fev database




```
##
## Call:
## lm(formula = logFev ~ logAge + logHt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62020 -0.08894  0.01166  0.09807  0.46645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.29520 osd: 0.39196 -26.266  < 2e-16 ***
## logAge        0.18045 ~0.85 0.03346   5.392 9.74e-08 ***
## logHt         2.62968 ~3.1 0.11010  23.884  < 2e-16 ***
## ---  Better?
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1481 on 651 degrees of freedom
## Multiple R-squared:  0.8033, Adjusted R-squared:  0.8026
## F-statistic:  1329 on 2 and 651 DF,  p-value: < 2.2e-16
```

R^2 for MLR ANOVA

Let's consider the coefficient of determination for MLR ANOVA, a.k.a. the "coefficient of **multiple** determination":



$$R^2 = \frac{\text{SSReg}}{\text{SST}} = \frac{\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}{\mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}$$

$\leftarrow$ = RSS if "$Y = \beta_0$"

It's not the square of correlation r anymore! Correlation is between two variables, whereas we have potentially many variables now.

However, as before, it's the proportion of the total sample variability in the Y's explained by the regression model.

Question: What happens to R^2 when you add more predictor variables? inc

$$R^2 = \frac{SST - RSS}{SST} = 1 - \frac{\overset{\downarrow}{RSS}}{SST}$$

RSS starts at SST, with 0 predictors.
RSS finishes at 0, with n-1 predictors

If R^2 goes from 0 to 1 in n-1 predictors, the avg contribution of a predictor is $\frac{1}{n-1}$

The effect on R^2 of additional predictors

Each time a predictor variable is added, SST stays the same because it depends on $\mathbf{Y}$ only.

However, adding a new predictor variable often improves (decreases) RSS: a richer model will often lead to a better fit, i.e. less error. Recall that

$$\text{RSS} = \hat{\mathbf{e}}'\hat{\mathbf{e}}$$

A least-squares minimization of RSS, with additional predictors now, is minimizing over a larger-dimensional space. This guarantees that the minimum is at least as small. So, at worst, RSS will stay the same (if you add a predictor that's ignored by fitting $\hat{\beta}_j = 0$), and usually it will get better.

If SST is constant and RSS decreases, SSReg must increase. Therefore R^2 will increase. (Put another way, the $\mathbf{H}$ in the numerator will have changed.)

Adjusted R^2

$$R^2 = 1 - \frac{RSS}{SST} = 1 - (n-p-1)\frac{MSE}{SST}$$

Because R^2 generally increases with the number of predictors, how do we compare the R^2 for a simple model to the R^2 for a many-variable model?

We can use the **Adjusted $\mathbf{R}^2$**, a better measure of the model fit. It is adjusted for the number of predictors in the model.

$$\text{Adj } R^2 = 1 - (n-1)\frac{\text{MSE}}{\text{SST}} = 1 - \frac{n-1}{n-p-1}\frac{\text{RSS}}{\text{SST}}$$

With additional predictor variables, the Adjusted R^2 will only increase if MSE decreases.

Exercise:

Show

$$F = \frac{R^2}{1-R^2}\,\frac{\overbrace{df_{denom}}^{n-p-1}}{\underbrace{df_{numer.}}_{p}}$$

Adjusted R^2 in action: First, reviewing regression ANOVA

For the `fev` vs age SLR dataset (PPC question 1), $n = 654$ and $p = 1$.

From Weeks 8–9 slide 22, $R^2 \approx 0.5722$ and Adj $R^2 \approx 0.5716 \approx R^2$, a difference of approximately only 0.1%.

Taking logs, and rerunning the analysis, today we got $R^2 \approx 0.6309$ and Adj $R^2 \approx 0.6303 \approx R^2$.

Adjusted R^2 in action: MLR ANOVA

Let's compare the (adjusted) coefficients of determination for a small dataset, with and without an extra predictor.

Consider just the first ten points in the `fev` database (A = abridged):

```
set.seed(1)
N<-10; u <- sample(length(logFev),N)
logFevA<-logFev[u]; logAgeA<-logAge[u]
rA<-rnorm(N) # A new potential predictor
                noise
mod2 = lm(logFevA~logAgeA)  <- without noise
mod3 = lm(logFevA~logAgeA+rA) <- with noise
summary(mod2) # SLR ANOVA
summary(mod3) # MLR ANOVA
```

Note that `rA` is noise, but adding it still increases the R^2. !

Results of SLR ANOVA

without noise

```
##
## Call:
## lm(formula = logFevA ~ logAgeA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34977 -0.04767 -0.00790  0.10280  0.26091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.6288     0.5944  -2.740  0.02544 *
## logAgeA       1.1232     0.2523   4.452  0.00213 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1747 on 8 degrees of freedom
## Multiple R-squared:  0.7125, Adjusted R-squared:  0.6765
## F-statistic: 19.82 on 1 and 8 DF,  p-value: 0.002132
```

Results of MLR ANOVA

with noise

```
##
## Call:
## lm(formula = logFevA ~ logAgeA + rA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32561 -0.05576 -0.01012  0.05902  0.29785
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.72678    0.64144  -2.692  0.03099 *
## logAgeA      1.16367    0.27176   4.282  0.00365 **
## rA           0.03408    0.05727   0.595  0.57055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1822 on 7 degrees of freedom
## Multiple R-squared:  0.7263, Adjusted R-squared:  0.6481
## F-statistic: 9.289 on 2 and 7 DF,  p-value: 0.01072
```

Next steps

- Assignment 2 was released on Wednesday 7 November. If you haven't received the Crowdmark email by now, check your spam folder
- Solutions to **PPC** were posted on 10 November
- Now that the exam date is known, additional TA office hours have been posted (click here)