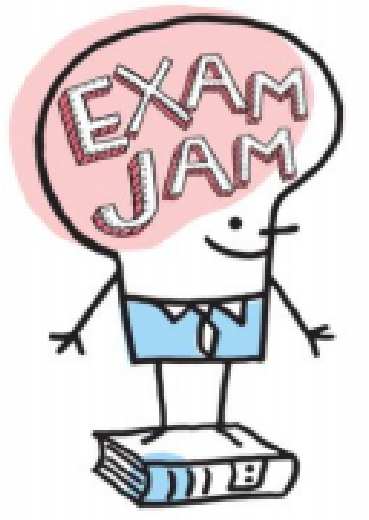
STA302/STA1001 Review/Study Session

Mark Ebden, 8 December

Up to slide 18 is new(ish), while slides 19–39 are mostly copy-and-paste

Welcome!



Thanks to the organizers:

http://www.artsci.utoronto.ca/current/exam_jam

"A **jam session** is a relatively informal musical event, process, or activity where musicians, typically instrumentalists, play improvised solos and vamp on tunes, songs and chord progressions." (Wikipedia)

This morning: After an exam overview and discussing certain requested topics, there will be time for small-group work.

Exam overview

The total number of marks available is 100. Therefore, there are 1.8 minutes available per mark (as with the midterm).

There are nine independent questions:

- Most questions have subparts
- The order of the questions doesn't follow the order in which we covered the course topics
- Topic coverage is as described in Portal under 'exam'

The questions

Question 1: Eight multiple choice questions. Instructions are: "Select one (1) letter per question, with no penalty for wrong answers. You don't need to show your work. Letters circled here won't be graded. Transfer your answers to your exam booklet."

Question 2: Explain/list/etc questions (not equation-based)

Question 3-4: Equation theory (proofs/models/calculations/etc)

Questions 5-8: Analysis aided by R code, graphs, etc

Question 9: Two definitions (pre-midterm)

Queries

"What are the questions like?"

- In between the style of the midterm and the style of old exams

"Are there proofs?"

- There are two proofs and one derivation

"How much R?"

- R appears in the multiple-choice section and among the four data-analysis questions
- It involves reading code (17 marks directly) and graph output (18 marks directly)

The first and last page are on Portal

UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER 2017 EXAMINATIONS

STA302H1S / STA1001H1S

Duration - 3 hours

Examination Aids: A calculator

First Name: _____ **Surname:** _____ **Student Number:** _____

Section: LEC0101/2001 (day) ☐ LEC0501 (evening) ☐

Instructions:

- Print your name on each exam booklet
- Use $\alpha = 0.05$ as a significance level
- **Hand in this Question Paper Please:** Do not enter any answers on this yellow question paper. It will be destroyed after the exam. Answers on this question paper will not be marked. All answers must be provided in the answer booklets included with your examination

Question	Value
1	16
2	13
3	8
4	16
5	6
6	13
7	10
8	12
9	6
Total	100

This exam should have 8 pages including this page

1

Topics requested by five students, in chronological order

- Two-sample t-test
- The midterm
- ANCOVA
- 7 C's for MLR
- Added Variable Plots
- Recent final exam
- Partial F-tests
- Interpreting MLR output about SSreg, MSreg etc

Today's Study Session topics

- **Two-sample t-test**
- ANCOVA
- Partial F-tests
- Added Variable Plots
- 7 C's for MLR
- Questions on the midterms or old exams

Taking a step back from our techniques

Remember this?

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

This represents some of the techniques we've discussed this term: SLR, DVR, the two-sample *t*-test, the *F*-test, MLR, and ANCOVA.

The reason for the mystery of the Trinity — the oneness among *t*-tests, *F*-tests, and dummy-variable regression — is the convergence of concepts in the above equation.

Comparing the techniques: Two kinds of regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

In our course, $\mathbf{Y}$ is a column vector of observations, and $\mathbf{e}$ is a column vector of errors meeting (at least) the Gauss-Markov conditions.

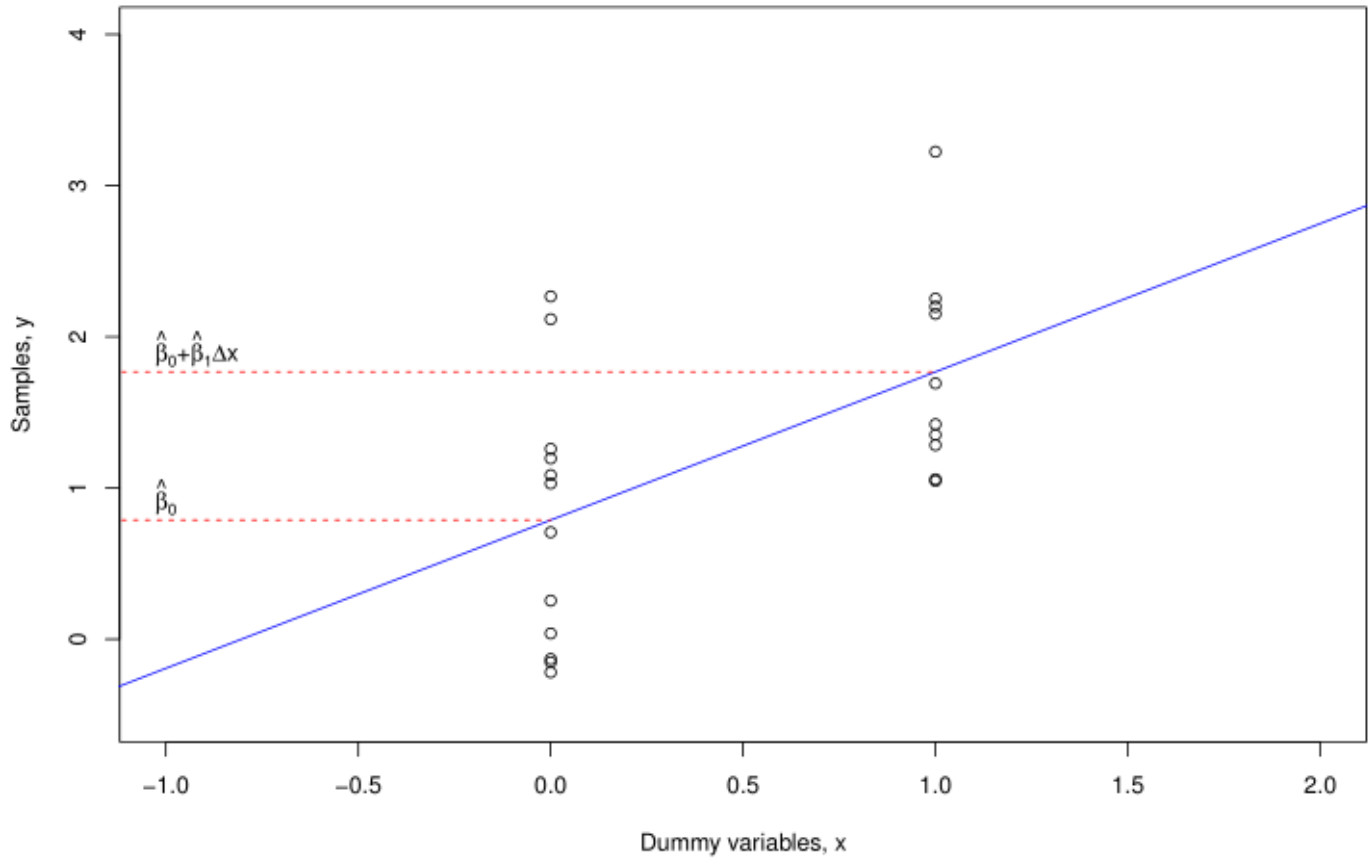
The $\mathbf{X}$ and $\boldsymbol{\beta}$ change depending on technique. For SLR/DVR they are:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

For SLR, the x_i's are continuous. The $\mathbf{X}$ are predictors and $\mathbf{Y}$ are response variables. Pairing all the (x_i, y_i) values makes a scatterplot. The β_0 and β_1 are the y-intercept and slope, respectively.

For DVR (dummy-variable regression), the x_i's are dummy variables. The $\mathbf{Y}$ are observations; each y_i comes from one of two sources (say A and B) as indicated by the corresponding $x_i = 0$ or 1.

Recall from Week 4: Dummy-variable regression



In Week 2 (slide 39 etc) we saw how least-squares methods fit to the mean (not median, etc) value of y for a given x. In DVR, the fitted line goes through both means, $\bar{y}_A$ and $\bar{y}_B$.

In our underlying model for DVR, the β_0 is the mean of the A group, μ_A. The β_1 is the difference between the two groups' population means, $\mu_B - \mu_A$.

Comparing the techniques: The two-sample t-test

Let's see if $\hat{\beta}_1$ on the previous slide is statistically significant. In our call to `lm`, R will calculate for us a test statistic (derived in Week 3):

$$T = \frac{\hat{\beta}_1 - 0}{\text{se}\left(\hat{\beta}_1\right)} \sim t_{n-2}$$

But since $\beta_1 = \mu_B - \mu_A$ and $\hat{\beta}_1 = \bar{y}_B - \bar{y}_A$, we can write

$$T = \frac{\bar{y}_B - \bar{y}_A}{\text{se}\left(\bar{y}_B - \bar{y}_A\right)} \sim t_{n-2}$$

This is exactly what a two-sample t-test does: once you calculate variance, the above becomes

$$\frac{\bar{y}_B - \bar{y}_A}{s\sqrt{1/n_A + 1/n_B}} \sim t_{n_A+n_B-2}$$

So, a two-sample t-test is equivalent to dummy-variable regression.

.

Comparing the techniques: the *F*-test

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

We've described $\mathbf{X}$ and $\boldsymbol{\beta}$ for SLR and for DVR. These two matrices grow fatter and longer, respectively, when we turn to ANOVA with $p + 1 > 2$ groups.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & & x_{2p} \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

The x_{ij}'s are dummy variables: 0's and 1's. Running least squares, again we fit through the mean of each group. Instead of a line hitting two means, we have a plane hitting three means or a hyper-plane hitting 4+ means. (Remember each predictor column in $\mathbf{X}$ gets its own dimensional axis.) The β_0 is the mean μ_0 of the reference group (indicated whenever a row in $\mathbf{X}$ has p zeros), and the other β_j values represent the difference $\mu_j - \mu_0$.

Comparing the techniques: the F-test

In STA302 we haven't concerned ourselves too much with the general problem of comparing the means of several groups, as is often done with ANOVA.

You'll recall though that the null hypothesis for the F-test (whether in ANOVA among groups or as applied to MLR as we did) is that $\beta_1 = \cdots = \beta_p = 0$. Namely, that the x_{ij} dummy variables aren't useful to predict the y_i's, i.e. the groups are similar – their population means are the same.

When using the ANOVA to compare the means of groups, we concerned ourselves mainly with a special case, i.e. when there are two groups ($p = 1$). Testing whether two population means are the same is what the two-sample t-test does, and we showed (Week 4, slide 11) that $T^2 = F$. You may have also done it as an exercise in Rice Chapter 6.

Comparing the techniques: MLR

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

For MLR with $p > 1$ predictors, the terms in the above equation at first look the same as for ANOVA. However, the x_{ij}'s are continuous and represent n observations each of p predictors, and the y_i's represent a response variable. You can produce multiple scatter plots from these. The β_0 is the y-intercept as with SLR, and the β_j's indicate how y changes when x_j changes alone.

The maths for the F-test are the same here, but $\mathbf{X}$ consists of continuous variables rather than dummy variables.

Comparing the techniques: ANCOVA

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

ANCOVA is when $\mathbf{X}$ contains a mixture of continuous variables and dummy variables. Remember there were a few possible ways to combine them:

- Parallel regression lines: $Y = \beta_0 + \beta_1 x + \beta_2 d + e$
- Regression lines with equal intercepts but different slopes: $Y = \beta_0 + \beta_1 x + \beta_3 d x + e$
- Unrelated regression lines: $Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 d x + e$

Advantages of using ANCOVA:

- We have tests for equal slopes and intercepts
- We have higher df_{error}, meaning the power increases and the CIs narrow
- We get a better estimate of σ^2 based on more observations

Possible disadvantage of using ANCOVA:

- An implicit assumption that both groups have the same error variance

Slide from Week 10: Application of ANCOVA

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 d x + e$$

We'll test whether the resulting change in Y (`Flowers`) when x (`Intensity`) changes is the same for early- versus late timings ($d = 1$ or 0). In other words, $H_0 : \beta_3 = 0$.

This isn't the same as asking: "Is the relationship between Y and `Intensity` the same for early and late timings? Do they have the same line?" (What is the hypothesis test in that case?)

The R code for our test is:

```
myFit <- lm(Flowers ~ Intensity * as.factor(Time), data=case0901)
summary(myFit)
```

R code overview for $\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$

```
y <- c(0,2,2,4,5,5,6); x1 <- (1:7); x4 <- rnorm(7)
x2 <- c(0,0,0,0,1,1,1)
x3 <- c(0,0,1,1,0,0,0)

lm(y~x1) # SLR for x-y relationship
lm(y~x1+x4) # MLR
anova(lm(y~x1)) # SLR ANOVA (F-test)
anova(lm(y~x1+x4)) # MLR ANOVA

lm(y~x2) # DVR for comparing 2 groups
t.test(y[1:4],y[5:7],var.equal=T) # t-test on 2 groups
lm(y~x1*x2) # ANCOVA

# Not a focus of STA302:
anova(lm(y~x2)) # ANOVA with 2 groups (incl F-test)
anova(lm(y~x2+x3)) # ANOVA with 3 groups
```

Study Session topics

- Two-sample t-test
- ANCOVA
- **Partial F-tests**
- Added Variable Plots
- 7 C's for MLR
- Questions on the midterms or old exams

Review of the partial F-test

This is useful when we want to test $H_0 : \beta_i = \cdots = \beta_k = 0$ versus H_a : at least one of $\beta_i, \ldots \beta_k$ isn't zero. This is for some subset of the p β's. The approach is:

1. Fit the model with all predictor variables (known as the *full model*), and calculate RSS, known as RSS(full)
2. Fit the model without the predictor variables whose coefficients we're testing (known as the *reduced model*), and calculate RSS, known as RSS(reduced)
3. Calculate the observed F:

$$F = \frac{(\text{RSS(reduced)} - \text{RSS(full)}) \,/\, (\text{df}_{\text{reduced}} - \text{df}_{\text{full}})}{\text{RSS(full)} \,/\, \text{df}_{\text{full}}}$$

The partial F-test

We know that:

- RSS in reduced model $\geq$ RSS in full model
- SSReg in reduced model $\leq$ SSReg in full model
- SST in reduced model $=$ SST in full model

Note that df_{full} is the number of degrees of freedom in the error for the full model. The difference $\text{df}_{\text{reduced}} - \text{df}_{\text{full}}$ is the number of parameters that you're testing in the partial F-test.

It can be shown that, under H_0, F_{obs} has an F distribution with ($\text{df}_{\text{reduced}} - \text{df}_{\text{full}}$, df_{full}) degrees of freedom.

The **intuition** behind the test is: Did RSS go down by a statistically significant amount when new predictors were added to the model?

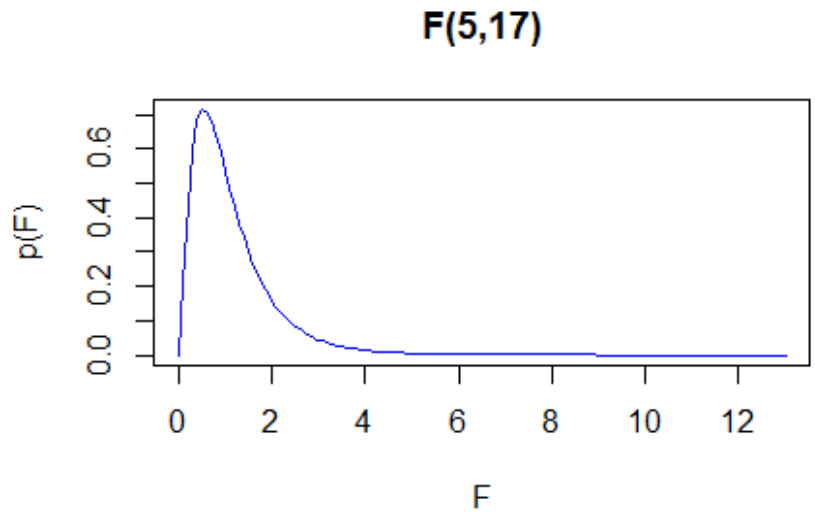
- Equivalently: Did R^2 increase by a statistically significant amount?

Recall from Weeks 10–11: "Is intensity important?"

$H_0 : \beta_1 = \cdots = \beta_5 = 0$ versus H_a : at least one of $\beta_1, \ldots \beta_5$ isn't zero.

We obtain a test statistic of

$$F_{\text{obs}} \approx \frac{(3451 - 767)/5}{767/17} \approx 11.9$$



There is strong evidence that not all of $\beta_1, \ldots \beta_5$ are zero, given that time is in the model. So we have reconfirmed that **yes** intensity is important.

The ANOVA table for the Meadowfoam dataset

We have decomposed SSReg into two components: intensity and timing.

Source	df	SS	MS	F
Regr(timing)	1	887	887	$887/45.15 = 19.6$
Regr(intensity)	5	2684	538	$538/45.15 = 11.9$
Error	17	767	45.15	
Total	23	4338		

Note that $887/45.15 \approx 19.6 \approx (4.43)^2$.

Also note that we could carry out a partial F-test on $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, i.e. on one parameter. Of course, this assumes all other variables are in the model.

Exercise: Try this for β_5, i.e. for $i750$'s coefficient.

Study Session topics

- Two-sample t-test
- ANCOVA
- Partial F-tests
- **Added Variable Plots**
- 7 C's for MLR
- Questions on the midterms or old exams

MLR Check 4: Added variable plots

An **added variable plot** allows you to *visualize* the relationship between a response variable and an explanatory variable over and above the other explanatory variables.

Such plots are also known as *partial regression plots*, *adjusted variable plots*, or *partial residual plots*.

The technique is based on the concept of **partial correlation**. The partial correlation between X_1 and X_2 given a set of n other variables $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_n\}$ is the correlation between:

- The residuals $\hat{e}_{X_1}$ resulting from the linear regression of X_1 versus $\mathbf{Z}$
- The residuals $\hat{e}_{X_2}$ resulting from the linear regression of X_2 versus $\mathbf{Z}$

Added variable plots

For the jth added variable plot, we first divide X into X_j and the other X's (excluding X_j). Then we plot:

- x-axis: Residuals from the regression of X_j versus the other X's
- y-axis: Residuals from the regression of Y versus the other X's

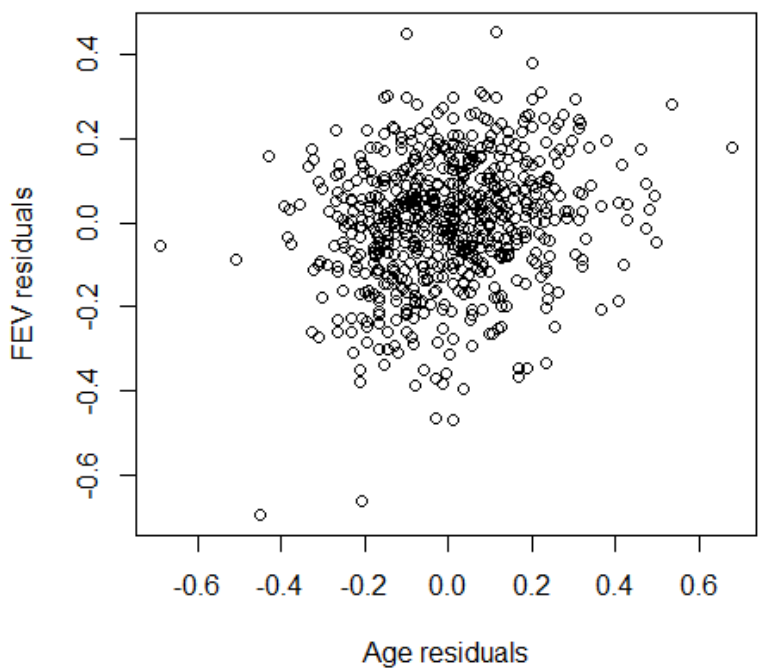
The correlation between the x- and y-axes here is a famous example of a partial correlation.

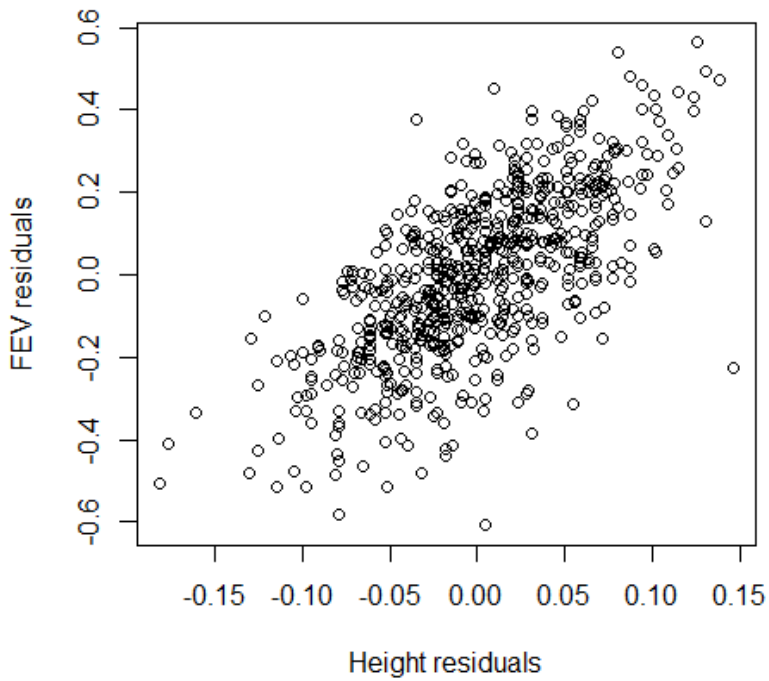
In an added variable plot, a linear pattern indicates that X_j is useful in the model over and above the other explanatory variables. In other words, the plot shows the strength of linear relationship Y and X_j over and above other variables.

The plot is also useful for detecting nonlinear relationships (polynomial in X_j for example), outliers, nonconstant variance, and influential points.

Example: FEV data

```
par(mfrow=c(1,2))
xAxis <- lm(logAge ~ logHt); yAxis <- lm(logFev ~ logHt)
plot(xAxis$residuals,yAxis$residuals,xlab="Age residuals",
     ylab="FEV residuals")
xAxis <- lm(logHt ~ logAge); yAxis <- lm(logFev ~ logAge)
plot(xAxis$residuals,yAxis$residuals,xlab="Height residuals",
     ylab="FEV residuals")
```





Study Session topics

- Two-sample t-test
- ANCOVA
- Partial F-tests
- Added Variable Plots
- **7 C's for MLR**
- Questions on the midterms or old exams

The Seven C's (7 Checks) for STA302 MLR diagnostics

1. Plot *standardized residuals* to help determine whether the proposed regression model is a valid model

2. Identify any *leverage points*

3. Identify any *outliers*

4. Assess the effect of each X on Y

5. Assess *multicollinearity*

6. Assess the assumption of *error homoscedasticity*

7. For time series: examine whether the data are *correlated over time*

Five traditional C's, now applied to MLR

1. The ith standardized residual is, as before,

$$r_i = \frac{\hat{e}_i}{S\sqrt{1 - h_{ii}}}$$

However, $S = \sqrt{\text{RSS}/(n - p - 1)}$ is the MLR estimate of σ; more important, there are superior techniques available, beyond the scope of this course.

2. To find leverage, we compute $\mathbf{H}$ as per earlier lectures. The threshold is:

$$h_{ii} > \frac{2(p + 1)}{n}$$

3. As before, outliers have $|r_i| > 2$ depending on the size of the dataset.

6. The constancy of variance is checked in a new way for MLR (see Check 1).

7. Correlations over time can be assessed as they were for SLR.

What about the two missing C's?

In MLR, we should still assess the assumption of **normal errors**: for this course, we can use the normal quantile plot of residuals as with SLR.

We can also still identify any **influential points**. The influence statistics are the same as those for simple linear regression:

$$\text{DFBETA}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\text{s.e. of } \hat{\beta}_{k(i)}} \qquad \text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\text{s.e. of } \hat{y}_{i(i)}}$$

$$D_i = \frac{\sum_{j=1}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} = \frac{r_i^2 h_{ii}}{2(1-h_{ii})}$$

However, the threshold formulae are generalized for MLR as follows:

- DFBETAS $> 2/\sqrt{n}$
- DFFITS $> 2\sqrt{\frac{p+1}{n}}$
- Cook's distance $D > \frac{4}{n-(p+1)}$

In each case, some authors check for a large gap as well, just as with SLR

Focus on MLR Check 1: Residual plots

Plotting (standardized) residuals versus X_j for $j \in \{1, \ldots p\}$ helps us to look for:

- Curvature
- Influential points
- Outliers

Plotting (standardized) residuals versus Y helps us to look for:

- Nonconstant variance
- Outliers

Plotting residuals versus other potential predictors can help expand our model as appropriate, as mentioned in our SLR work.

And as with SLR, residual plots are not the only way to assess model assumptions. For example, plots of Y vs X_j help us answer:

- Is the linear model appropriate?
- Are there unusual points? (e.g. potential outliers or influential points)

We can also look at the added variable plots (Check 4, to come).

MLR Check 5: Multicollinearity

Multicollinearity occurs when there is lots of correlation among the X's. We use the term interchangeably with "ill-conditioning".

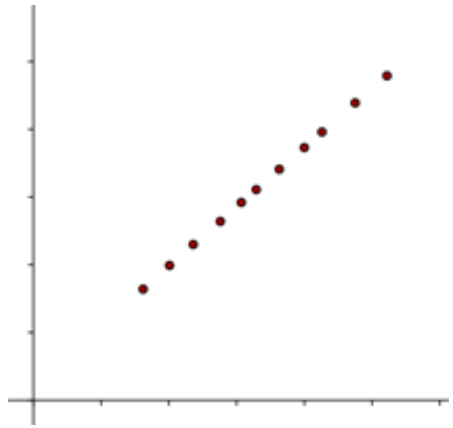
When explanatory variables are highly correlated, it's difficult or impossible to measure the individual variable's influence on the response.

The fitted equation is unstable:

- The estimated regression coefficients vary widely from data set to data set (even if the data sets are very similar), and depending on which other predictor variables are in the model
- An estimated coefficient may have opposite sign to what you'd expect
- A coefficient might not be statistically significantly different from zero even though there is a strong relationship between the X and Y when only considering X and Y

Multicollinearity

Recall: $\hat{\beta} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. If some the X's are perfectly correlated, $\mathbf{X}'\mathbf{X}$ is singular and we can't calculate $\mathbf{b}$.



Put another way, in terms of the Week 8 SLR material: if X contains linearly dependent columns, X has a rank below $p+1$. Therefore $(\mathbf{X}'\mathbf{X})^{-1}$ has a rank below $p+1$. A matrix must have full rank to be invertible.

In the case of $\mathbf{X}'\mathbf{X}$ *close* to singular, the determinant of $\mathbf{X}'\mathbf{X}$ will be *near* 0. Therefore, $\text{var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ will be large. This means that the standard errors of the estimated coefficients will be large, so we'll have "inefficient" estimates. (We can't make precise statements about their values.)

Quantifying Multicollinearity

Let R_j^2 represent the coefficient of multiple determination obtained when the jth predictor variable is regressed against the other predictor variables.
The **variance inflation factor** is

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

A large VIF_j is a sign of multicollinearity. Rules of thumb:

- If $5 \lesssim \text{VIF}_j \lesssim 10$, the effects of multicollinearity might be seen (this is a warning)
- If $\text{VIF}_j \gtrsim 10$, there is a serious problem

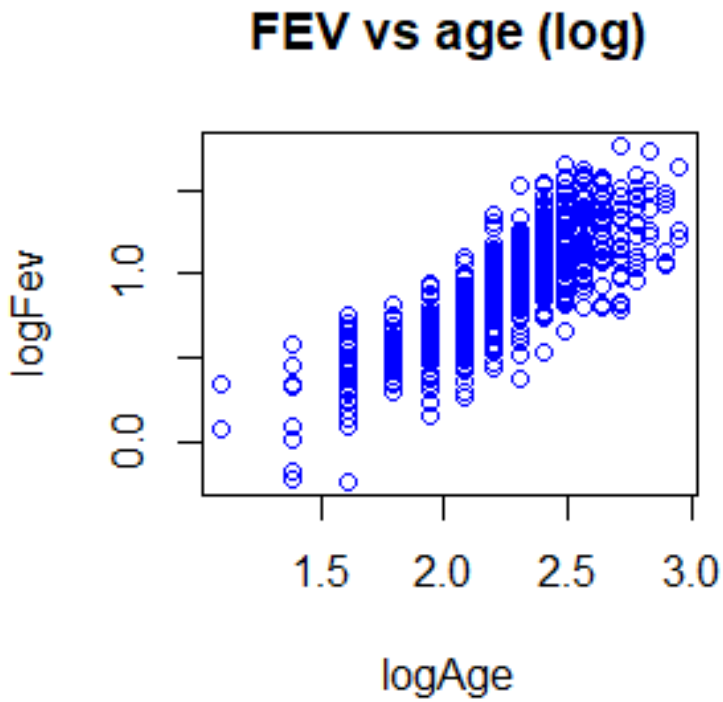
However, if the jth variable's coefficient is statistically significantly different from zero, this can be an indication not to worry as much about a high VIF_j.

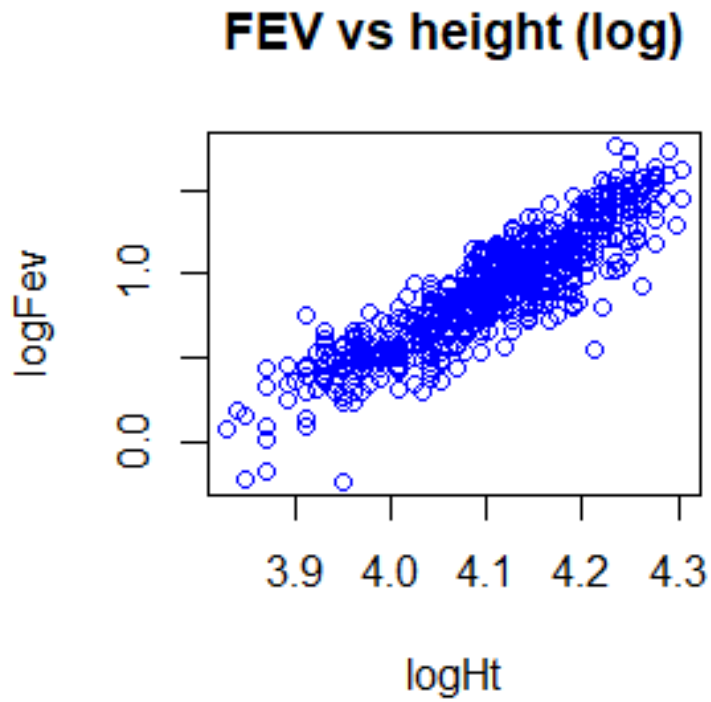
Optional material: **Tolerance** is defined as $1/\text{VIF}_j$

Multicollinearity in the FEV dataset

We'd calculated (on slides 10 and 20) that $\beta_1 \approx 0.18$ for the relationship between `logFEV` and `logAge` in a particular MLR context.

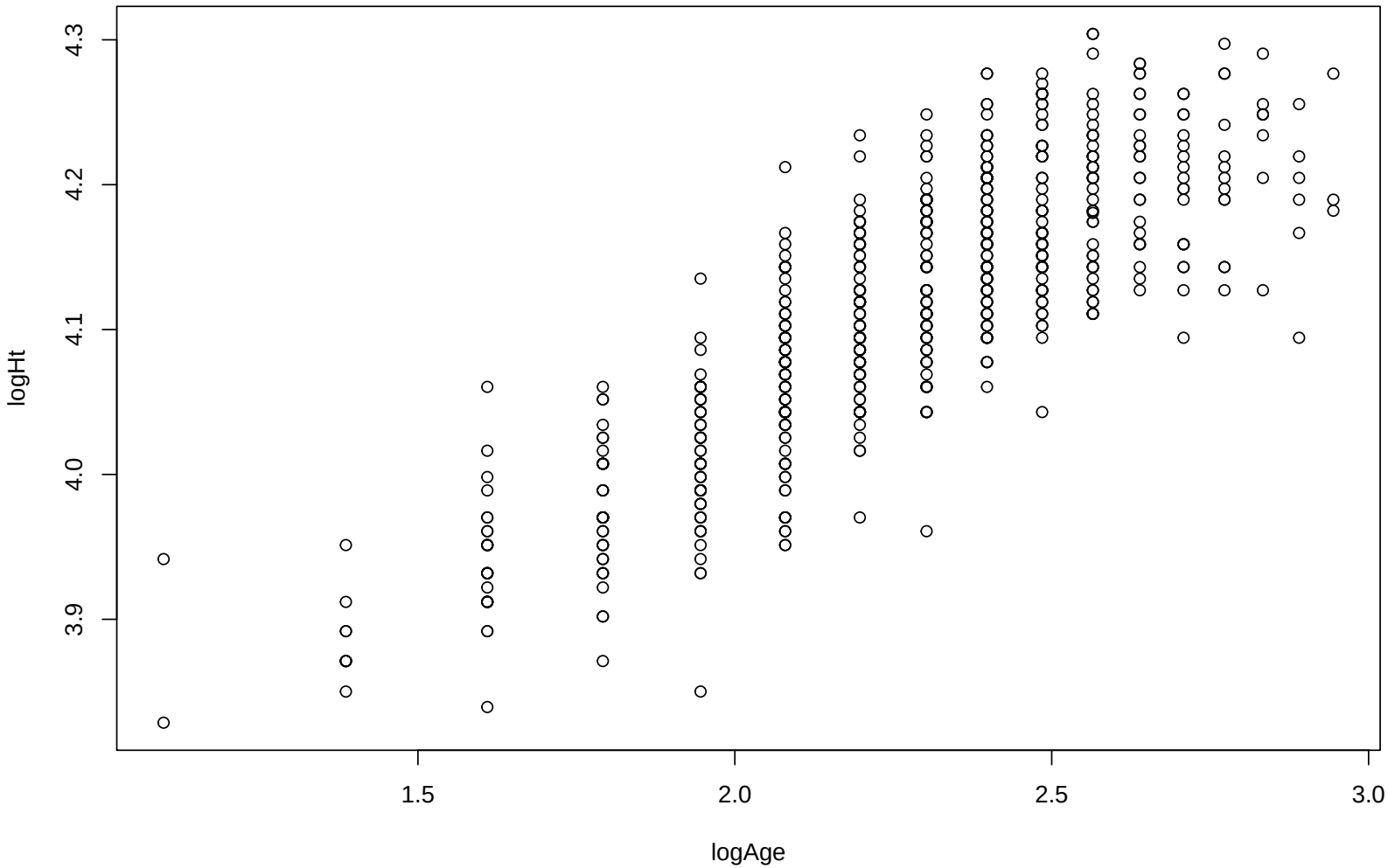
Here are the graphs from slide 7:





For this dataset, $\text{VIF}_1 = \text{VIF}_2 \approx 3.3$.

Scatterplot of the predictor variables

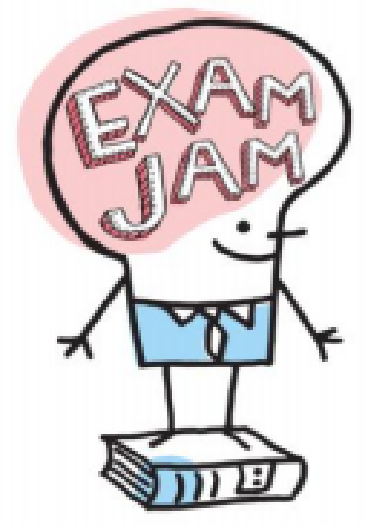


Study Session topics

- Two-sample t-test
- ANCOVA
- Partial F-tests
- Added Variable Plots
- 7 C's for MLR
- **Questions on the midterms or old exams**

To be handled at the same time as small-group work.

I think, therefore I jam



Small-group work:

- Suggested group size of not more than four
- Topic for most groups: "What question is bugging **me** from the midterms or from an old exam?"
- A special group will have the topic of R (including homework help)
- If you aren't in a Recognized Study Group, please meet somebody new!

In addition to our session

Until 3 pm in the Sid Smith lobby there is: massage therapy, button making, a Plinko game by APUS, puppies, vegan snacks and recipes from the Sustainability Office, the Innovation Hub's Chill Spot, learning to make balloon animals, blender-bike smoothies, a photo booth, free snacks and hot drinks, and free giveaways.