Math 455 Class 4 January 29

Hints on homework:

(1) The universal textbook problem-solver:

to solve the exercises in Section 8.5,

look at the examples and results in Section 8.5.

(2) Virtually every exercise in Wackerly, et. al.

is solved online.

Warning: make sure you can produce solutions yourself,

because you'll have to do that for the tests.

(3) If all else fails, you could come talk to me.

Q: What is the point of all these strange exercises

where we determine the distribution of some random

algebraic combination of random variables?

A: In the regression output, we produce statistics

which turn out to have the t- and F-distributions.

None of the hypotheses of regression mention t-

or F-distribution. Why do the statistics have

these distributions?

Q: How do we show that <statistic> has <distribution>?

A: Refer to definition of <statistic> and <distribution>!

Math 455 Class 04 January 2~~4~~9

Homework: hints and suggestions

Q: The first homework problem in HW 2 asks "What is the definition of $F_{0.01}$?"

What kind of answer is expected?

A: The hint says: see appendix on F-dist. On p. 852 there is a diagram defining F_α for any α. The homework asks you to translate this into words. Your answer should say:

" $F_{0.01}$ is the real number such that

Probability (Something involving $F_{0.01}$)

= something "

Figure out what the somethings are by

looking at diagram p. 852.

Note that this $F_{0.01}$ will depend on v_1 (the numerator dof) and v_2 (the denominator dof)

Similar remarks hold for the problem asking for a definition of $\chi^2_{0.01}$.

Q: In problems like 7.37 and 7.38, when we are asked "why?" what sort of answer is expected?

A: We'll look at 7.38 (a).

First observe that
$$\frac{\sqrt{5}\, Y_6}{\sqrt{W}} = \frac{Y_6}{\sqrt{W/5}}$$

Next note that Y_6 is std normal, because this is a hypothesis of 7.37.

Now $W = Y_1^2 + Y_2^2 + \cdots + Y_5^2$ defined in (7.37a)

and Y_i are indep standard normal RVs.

So $W \sim \chi^2$ with 5 dof by Thm. $\overset{6.4}{\cancel{6.10}}$ p. 321.

Also W and Y_6 are independent, because

W depends only on $Y_1, \ldots, Y_5$ and

$Y_1, \ldots, Y_5, Y_6$ are indep because this is a hypothesis of 7.37

Thus $\dfrac{Y_6}{\sqrt{W/5}}$ has the t-dist with 5 dof

by definition 7.2.

Q: Why is this relevant? To regression?

A: The regression model assumes that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where β_0, β_1 are fixed but unknown constants and $\varepsilon_i \sim N(0, \sigma^2)$ are independent.

We are not able to observe the values of ε_i

We only see our data (X_i, Y_i)

We construct statistics (functions of our data) which then implicitly depend on the ε_i (since Y_i does)

There are theorems which say <statistic> has a t-dist / F-dist < dof info >

These theorems are proved by noting that the <statistic> is a function of normal RVs and then doing stuff much like the solution to 7.38 (a) only more complex.

Heuristic: if you're summing squares of normals, expect χ^2 (Thm. 6.$\overset{4}{1}$0)

If you have $\dfrac{\text{normal}}{\langle \text{a std. dev} \rangle}$

expect t-dist. (Def 7.2).

If you have $\dfrac{\text{sum of sq}}{\text{sum of sq.}}$

expect F-dist. (Def. 7.3)

Recall the Universal Textbook Problem-Solving Method: If the problem is in Section 7.2 the solution probably involves ideas from the def'ns, theorems and ~~exa~~ examples of 7.2.

Regression analysis is fundamentally about drawing a line through a bunch of points (which don't all lie on a line)

We have data (X_i, Y_i) $i = 1, \ldots, n$

We want to draw a line $Y = \beta_0 + \beta_1 X$ which "best" fits these points.

"best" means: minimize the sum of squared errors, Q

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2$$

Notice that the "errors" are differences in Y-coordinate.

This means that the X-variable and Y-variable are being treated differently.

This means in Ch. 7 errors in measurement of X and errors in measurement of Y have different consequences.

So, suppose that we accept that minimizing SSE Q is "best". What is the best β_0 (call it $\hat{\beta}_0$) and best β_1 (call it $\hat{\beta}_1$)?

A: Calculus: set $\dfrac{\partial Q}{\partial \beta_0} = \dfrac{\partial Q}{\partial \beta_1} = 0$.

get "normal equations" for $\hat{\beta}_0, \hat{\beta}_1$

$$n\beta_0 + \beta_1 \Sigma X_i = \Sigma Y_i \qquad \left(\Sigma = \sum_{i=1}^{n} \right)$$

(*)

$$\beta_0 \Sigma X_i + \beta_1 \Sigma X_i^2 = \Sigma X_i Y_i$$

Writing $\bar{X} = \frac{1}{n} \Sigma X_i$ $\quad \bar{Y} = \frac{1}{n} \Sigma Y_i$

we have $\hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y}$

This means that $(\bar{X}, \bar{Y})$ is always on the regression line $\hat{\beta}_0 + \hat{\beta}_1 X = Y$.

More work with the 2 normal equations $(*)$ gives

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

(Think: $\hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)}$)

Technical point: how do we know that

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = 0$$ gives us a min,

and not a max or saddle point?

A: The "Second Derivative Test" from Calc. 3.

Check that $\frac{\partial^2 Q}{\partial \beta_0^2} > 0$ and that

$$\frac{\partial^2 Q}{\partial \beta_0^2} \frac{\partial^2 Q}{\partial \beta_1^2} - \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} > 0.$$

[Think: $f_{xx} > 0$ and $f_{xx} f_{yy} - f_{xy} > 0$]

This will always be satisfied unless there is some degenerate condition, e.g. all data are on a line, or there ~~is~~ are no data, etc.

Algebraic tricks for getting $\hat{\beta}_1$ into the form given above are covered in Wackerly, Ch 1, Ex. 1.13.

Example of such a trick: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

so $\sum X_i = n\bar{X}$, and $\sum (X_i - \bar{x})$

$= \sum X_i - \sum \bar{x} = n\bar{X} - n\bar{X} = 0$

Note that the homework requires you to use R at certain points, e.g. finding $F_{0.01}$ for various values of ν_1, ν_2.

The solution here is "help(df)" then figure out which of df, pf, qf etc you need to use and which arguments you need to give the R function.

~~Proposed date for Midterm 1:~~

~~Wednesday before Spring Break.~~