73-360, Spring 2000
Final Exam, Solution

The SAS output is displayed below. Only the relevant parts have been retained. Models 1, 2 and 3 on pages 8, 9 and 10 of the original SAS output have been renamed as models 5, 6 and 7. Pages 6 and 7 of the SAS output are irrelevant for us and have been removed.

Also, I have only answered the questions which are in the scope of this class.

The Maximum Likelihood Estimates are precisely the Least Squares estimates, since we studied in class that LS is the Best Linear Unbiased Estimator under various assumptions which we expect to be satisfied here.

Variable	N	Mean	Std Dev	Minimum	Maximum
INCOME	5321	128.0013155	89.5245346	0	975.0000000
AGE	2929	1933.99	2.2354895	1924.00	1938.00
PCTHMO	2287	13.0524705	21.1077072	0	100.0000000
EXPER	2929	2.5380676	1.1082166	1.0000000	4.0000000
HOURS	5820	58.3848797	18.2964451	1.0000000	190.0000000
GEND	6053	0.2592103	0.4382374	0	1.0000000
WAGE	5113	47.7812468	45.6926384	0	1800.00
GENERAL	6053	0.4733190	0.4993289	0	1.0000000

Model: MODEL1
Dependent Variable: INCOME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	2014953.7416	503738.4354	93.853	0.0001

Error	2536	13611517.662	5367.31769
C Total	2540	15626471.403	

Root MSE	73.26198	R-square	0.1289
Dep Mean	114.56828	Adj R-sq	0.1276
C.V.	63.94613		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-5805.853932	1358.3586749	-4.274	0.0001
GENERAL	1	-33.546498	2.99880322	-11.187	0.0001
EXPER	1	7.435558	1.38345610	5.375	0.0001
AGE	1	3.065602	0.70275673	4.362	0.0001
GEND	1	-34.771403	3.25490915	-10.683	0.0001

Model: MODEL2
Dependent Variable: HOURS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	83952.42777	20988.10694	66.010	0.0001
Error	2536	806331.33925	317.95400		
C Total	2540	890283.76702			

Root MSE	17.83126	R-square	0.0943
Dep Mean	58.41244	Adj R-sq	0.0929
C.V.	30.52649		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-1309.389923	330.61149917	-3.961	0.0001
GENERAL	1	2.186953	0.72988000	2.996	0.0028
EXPER	1	-0.857029	0.33671997	-2.545	0.0110
AGE	1	0.709611	0.17104426	4.149	0.0001

GEND	1	-12.255900	0.79221373	-15.470	0.0001

Model: MODEL3
Dependent Variable: WAGE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	4	137977.49151	34494.37288	23.721	0.0001
Error	2536	3687845.3025	1454.1976745		
C Total	2540	3825822.794			

Root MSE	38.13394	R-square	0.0361
Dep Mean	42.10631	Adj R-sq	0.0345
C.V.	90.56585		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-738.926772	707.04577689	-1.045	0.2961
GENERAL	1	-12.536229	1.56092142	-8.031	0.0001
EXPER	1	3.238045	0.72010936	4.497	0.0001
AGE	1	0.403050	0.36579527	1.102	0.2706
GEND	1	-0.768964	1.69422834	-0.454	0.6500

Model: MODEL4
Dependent Variable: WAGE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	135838.97440	67919.48720	46.716	0.0001
Error	2538	3689983.8196	1453.894334		
C Total	2540	3825822.794			

Root MSE	38.12997	R-square	0.0355
Dep Mean	42.10631	Adj R-sq	0.0347
C.V.	90.55641		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	39.916820	1.99315784	20.027	0.0001
GENERAL	1	-12.965354	1.51695034	-8.547	0.0001
EXPER	1	3.492674	0.68456763	5.102	0.0001

Model: MODEL5
Dependent Variable: INCOME

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	2056440.0251	293777.14644	54.837	0.0001
Error	2533	13570031.378	5357.2962409		
C Total	2540	15626471.403			

Root MSE	73.19355	R-square	0.1316
Dep Mean	114.56828	Adj R-sq	0.1292
C.V.	63.88640		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-6424.128442	1973.3047163	-3.256	0.0011
GENERAL	1	982.269384	2719.8337095	0.361	0.7180
EXPER	1	8.284509	1.97619288	4.192	0.0001
AGE	1	3.385279	1.02099193	3.316	0.0009
GEND	1	-44.758694	5.04646571	-8.869	0.0001
GEXP	1	-1.902806	2.76767747	-0.688	0.4918
GAGE	1	-0.525148	1.40739387	-0.373	0.7091
GGEN	1	17.043641	6.60285828	2.581	0.0099

Model: MODEL6
Dependent Variable: HOURS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	84631.97977	12090.28282	38.012	0.0001
Error	2533	805651.78725	318.06229		
C Total	2540	890283.76702			

Root MSE	17.83430	R-square	0.0951
Dep Mean	58.41244	Adj R-sq	0.0926
C.V.	30.53169		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-1589.781937	480.81434398	-3.306	0.0010
GENERAL	1	503.018334	662.71318868	0.759	0.4479
EXPER	1	-1.289571	0.48151807	-2.678	0.0075
AGE	1	0.855190	0.24877433	3.438	0.0006
GEND	1	-12.935260	1.22961907	-10.520	0.0001
GEXP	1	0.827326	0.67437077	1.227	0.2200
GAGE	1	-0.260191	0.34292482	-0.759	0.4481
GGEN	1	1.109928	1.60884882	0.690	0.4903

Model: MODEL7
Dependent Variable: WAGE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	141981.51093	20283.07299	13.947	0.0001

```
Error          2533 3683841.2831 1454.3392353
C Total        2540  3825822.794

    Root MSE       38.13580     R-square       0.0371
    Dep Mean       42.10631     Adj R-sq       0.0345
    C.V.           90.57026

                        Parameter Estimates

                 Parameter      Standard    T for H0:
Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

INTERCEP   1   -919.378593  1028.1445629        -0.894        0.3713
GENERAL    1    304.779744  1417.1061455         0.215        0.8297
EXPER      1      3.789753    1.02964937         3.681        0.0002
AGE        1      0.495941    0.53196412         0.932        0.3513
GEND       1     -3.438693    2.62934368        -1.308        0.1911
GEXP       1     -1.155928    1.44203403        -0.802        0.4229
GAGE       1     -0.163207    0.73328987        -0.223        0.8239
GGEN       1      4.565157    3.44026586         1.327        0.1846
```

1. Holding constant experience, age, and gender, how much more do specialists make than generalists, at your best guess? 95% confidence interval? How many more hours do specialists work (confidence interval)?

   MODEL1 regresses income as a function of generalists, experience, age and gender. The coefficient of GENERAL is -33.55, indicating that our best guess is that a generalist makes \$33,550 less per year than a specialist. A 95% confidence interval is computed in the usual way, using the standard error computed as 2.999, and $z_{0.025} = 1.96$. We are using the normal distribution because we have sufficiently many observations (2540, as noted under the DF coloumn of any of the models. Hence a 95% confidence interval is $33550 \pm 2999 \times 1.96 = [27672,\ 39428]$.

   MODEL2 regresses hours on the same four variables. Again, the coefficient and standard error of General answers the question. Our best estimate is that holding all else constant, a generalist works 2.187 hours more per week than a specialist. A 95% confidence interval is computed as above to be $[0.756,\ 3.618]$.

2. A critic claims that your results are wrong. She claims that more experienced doctors have incomes, wages, and hours that are more "spread out" than do less experienced doctors and that your results are worthless because they do not account for this. How would you respond?

   This critic is saying that there is heteroskedasticity in the regression: that the variance of the errors (hence the variance of the LHS variables incomes, wages, hours) are higher for more experienced doctors. We should respond that this problem, if it exists, will NOT bias our coefficient estimates but will bias our standard errors. We could do the estimation with robust standard errors to respond to the critic.

3. Does the difference in income between generalists and specialists come mostly from a difference in hours, a difference in wages, or a combination?

   We first note that in MODEL2, the p-value of GENERAL is less than 1%. Hence we have significant statistical evidence to conclude that generalists do in fact work more than specialists.

   In MODEL3, the p-value of GENERAL is again much less than 1%. Hence we also conclude that we have sufficient statistical evidence that the hourly wages of generalists is less than that of specialists.

   Hence despite working longer hours, generalists earn much less than specialists. We conclude that among the two reasons given, the fact that generalists have a much lower hourly wage is the more important reason why generalists have a much lower annual income than specialists.

4. One theory of why female MDs make less income than do male MDs is that females typically have child-care obligations which result in them working fewer hours than do males. What evidence (either for, against, or both) regarding this theory can you find in the results? Please be thorough in your answer.

   Again, we know that annual income is a function of hours worked per week as well as hourly wage. We first look at MODEL2 and conclude from the low p-value of GEND that females work fewer hours than males. It is beyond our purview to deduce whether or not the fewer hours are due to child care obligations.

   MODEL3 has WAGE as the dependent variable, and we note that the p-value of GEND is very high at 65%. Hence we would almost certainly accept the null hypothesis that the population coefficient of GENDER is zero. That is, we find no evidence to support the hypothesis that female doctors have a different hourly wage than male doctors.

   Also, MODEL1 tells us that female doctors do in fact have a lower annual income (again, evidenced by the low p-value of GEND). Hence this lower annual income must be due to lower hours.

5. Do age and gender affect wages (holding constant type of doctor and experience)?

   Here we want to test the null hypothesis that the coefficients of both AGE and GEND are zero, against the alternative that at least one of them is non-zero. To do this, we use the F-test for subsets of variables.

   Our unrestricted model is MODEL3, which regresses wages on age, gender, experience and type of doctor. The sum of squared errors is $SSE_{UR} = 3687845$, which appears under `Sum of squares` corresponding to `Error`.

   The restricted model is MODEL4, which regresses wages only on type of doctor and experience. Here we find $SSE_R = 3689984$.

   Our F-statistic is $\frac{(SSE_R - SSE_{UR})/2}{SSE_{UR}/(2540-4-1)} = \frac{1069.5}{1454.8} = 0.735$. Since this is lower than $F_{2,2535,0.05} = 3.00$, we accept the null hypothesis. We conclude that we have sufficient statistical evidence (95% confident) that neither age nor gender impact wages.

7. A critic of your results above claims that you are assuming that the effects of age, experience, and gender on income, wages and hours are the same for generalists and specialists. Is this true? If you can, test to see if that assumption is valid.

In effect, here we want to test whether we can "pool" generalists and specialists in the same model, for INCOME, WAGES and HOURS. Here we do the test only for INCOME; the other two can be tested in the same way.

Our null hypothesis is that the coefficients are the same for EXPER, AGE and GEND for generalists and specialists; or in other words, all three of the coefficients GEXP, GAGE, GGEN are zero.

For INCOME, the model which separates generalists from specialists is MODEL5, on page 8 of the SAS output. The last three variables with the prefix G are variables only for generalists. The sum of squared errors in the separated model is $SSE_{UR} = 13570031$.

The pooled model is the original MODEL1, where we do not separate coefficients for AGE, EXPER and GEND based on type of doctor. The sum of squared errors here is 13611517. In other words, this is the restricted model, where we are assuming that the effects of AGE, EXPER and GEND are the same for both types of doctors.

We form our F-statistic: $\frac{(SSE_R - SSE_{UR})/3}{SSE_{UR}/(2540-7-1)} = \frac{13829}{5359} = 2.58$. This is less than $F_{3,2532,0.05} = 2.60$, and so we accept the null hypothesis. We are able to refute the critic and say that we do have statistical evidence to support our assumption that the effects of age, experience and gender on income is the same for specialists and generalists.

Note that we just barely rejected the null, so the conclusion isn't all that strong.

We can test for the effects on WAGES and HOURS in a similar way, using the other models.

8. Looking at the model on page 10, tell me about (your best estimate of) the pattern of wages for females vs males in generalist and specialist disciplines (holding constant experience and age).

   For specialists, the difference between male and female doctors in terms of wages is given by the coefficient GEND = -3.43. That is, female specialists have an hourly wage which is \$3.43 lower than male specialists, holding all else constant, at our best guess. Noting the high p-value, we aren't very confident about this, and it may in fact be the case that there is no difference due to gender for specialists.

   For generalists, the difference between males and females is the sum of the coefficients GEND and GGEN. Therefore our best guess is that holding all else constant, female generalists have an hourly wage which is \$0.78 (by adding the coefficients of GEND and GGEN) more than male generalists. Again, the high p-value makes our conclusion suspect.