

Tools for Teaching Econometrics

Pierre Chausse*

Abstract

This vignette explains how to use the different tools I used in my econometrics courses at the University of Waterloo. It includes functions to generate solutions from inference questions, to print regression results and more.

1 Introduction

I use many functions in my courses to print regression results in a nice format, generate solutions for inference question on the mean, the variance, and least squares models, to simulate data to illustrate concepts, and so on. This document explains how to use them using examples from the courses. To run the different functions available in the package, you first need to load it:

```
library(metricsUW)
```

2 Printing regression results in equation format

To print a regression result from `lm` or `glm`, the function `printReg` generates a latex equation with the coefficient estimates and their standard errors. To have the equation printed in Latex format in a R-Markdown document, the chunk option `results='asis'` must be added. For example, the following is an estimated wage equation using the PSID1976 dataset from the `AER` package:

```
data(PSID1976, package="AER")
fit1 <- lm(wage~education+age+I(age^2), PSID1976)
printReg(fit1)
```

$$\widehat{wage} = -8.5590 + \frac{0.4538}{(0.0495)} education + \frac{0.2558}{(0.1524)} age - \frac{0.0029}{(0.0018)} I(age^2)$$
$$n = 753, \quad R^2 = 0.1047, \quad SSR = 7075.373$$

The function offers different options. You can add stars for significant coefficients (`stars=TRUE`), adjusted R^2 (`adjrsq=TRUE`), limit the number of variables per line when the equation does not fit the page (e.g. `maxpl=3`), replace default standard errors by robust ones (`se=newse`), replace the default t-distribution used to compute p-values by the $N(0,1)$ distribution (`dist="n"`) or omit variables. We present here a few examples.

- Adding adjusted R^2 and stars, and reducing the number of variables per line. I also increase the number of digits.

```
fit2 <- lm(wage~education*city+heducation+age+I(age^2), PSID1976)
printReg(fit2, maxpl=3, adjrs=TRUE, stars=TRUE, digits=5)
```

*University of Waterloo, pchause@uwaterloo.ca

$$\widehat{wage} = -6.83341 + 0.45828 \text{ education} - 1.42391 \text{ cityyes} - 0.12139 \text{ heducation}$$

$$+ 0.24075 \text{ age} - 0.00284 I(\text{age}^2) + 0.13427 \text{ education : cityyes}$$

$$n = 753, R^2 = 0.11471, SSR = 6996.553, \bar{R}^2 = 0.10759$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

- Replacing default standard errors by robust ones, and using the standard normal distribution for p-values.

```
library(sandwich)
newse <- sqrt(diag(vcovHC(fit2)))
printReg(fit2, maxpl=3, se=newse, stars=TRUE)
```

$$\widehat{wage} = -6.8334 + 0.4583 \text{ education} - 1.4239 \text{ cityyes} - 0.1214 \text{ heducation}$$

$$+ 0.2408 \text{ age} - 0.0028 I(\text{age}^2) + 0.1343 \text{ education : cityyes}$$

$$n = 753, R^2 = 0.1147, SSR = 6996.553 \text{ (Robust S-E)}$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

For GLM estimation, the R^2 is replaced by residual deviance and AIC is printed. Also, the left-hand side specifies that it is the link of \hat{Y} that has the linear representation. For example, the following is the result from a Poisson regression with the log link:

```
data(fertil2, package="wooldridge")
fit3 <- glm(children~educ+age+I(age^2)+catholic+electric+radio+tv+heduc,
            family=poisson(link=log), data=fertil2)
printReg(fit3, maxpl=3, stars=TRUE)
```

$$\text{link } [\widehat{\text{children}}] = -3.8200 - 0.0187 \text{ educ} + 0.2650 \text{ age} - 0.0031 I(\text{age}^2)$$

$$+ 0.0104 \text{ catholic} - 0.0803 \text{ electric} + 0.0444 \text{ radio}$$

$$- 0.1229 \text{ tv} - 0.0126 \text{ heduc}$$

$$n = 1953, AIC = 7344.376, \text{ Residual Deviance} = 1834.816,$$

$$\text{ Null Deviance} = 3246.903, \text{ Family} = \text{poisson}, \text{ Link} = \text{log}$$

*pv < 0.1; **pv < 0.05; ***pv < 0.01

If we just want to create a regression equation from a formula, we just set the argument `form` to the desired formula. Here is an example:

```
printReg(form=log(wage) ~ education*female+age+I(age^2))
```

$$\log(wage) = \beta_0 + \beta_1 \text{education} + \beta_2 \text{female} + \beta_3 \text{age} + \beta_4 I(\text{age}^2) + \beta_5 \text{education : female} + u$$

3 Solution to inference questions

The way the functions are organized is as follows. First the inference function generates the solution, which is an object of class `metricsSol`, and the `print` method generates the answer in Latex format. It is meant to be printed inside an R-Markdown chunk, with the option `results="asis"`.

3.1 Introduction to Statistics

This section covers inference problems typically covered in an introductory statistics course.

3.1.1 Test on the mean

The solution is based on the following properties:

- If $X_i \sim N(\mu, \sigma^2)$, then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \sim t_{n-1},$$

where n is the sample size, \bar{X} is the sample mean and $\hat{\sigma}$ is the sample standard errors defined as $\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$.

- If the distribution of X_i is unknown, the true distribution of the test is also unknown, but the distribution it converges to is known. In my course, the t-distribution is only used when it is the exact distribution. If not, we use the $N(0,1)$ as an approximation. The solution generator applies this rule as well.

Suppose we have a series and want to test the hypothesis $H_0 : \mu = c$ against $H_1 : \mu \neq c$, $H_1 : \mu > c$ or $H_1 : \mu < c$. The function `testm` generates the solution. Consider the `wage` series form the PSID1976 dataset, restricted to workers with positive hours:

```
wage <- subset(PSID1976, hours>0)$wage
```

We want to test if the population average is equal to 5 dollars per hour against the alternative that it is not equal to 5. We just insert the following code in the chunk:

```
testMean(wage, h0=5)
```

Testing $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ at 5%

$$test = \frac{(\bar{x} - 5)}{s/\sqrt{n}} = \frac{(4.1777 - 5)}{(3.3103)/\sqrt{428}} = -5.1392 \sim t_{427}$$

Since $|-5.1392| > 1.9655$ (the 97.5% quantile of the t_{427}), we reject H_0 .

It is not necessary to use the `print` method directly, unless we want to add something to the solution. For example, I like to add mark distribution in my exam solutions:

```
sol1 <- testMean(wage, h0=5)
print(sol1, addMess="1 point for the statistic and 1 point for the conclusion")
```

Testing $H_0 : \mu = 5$ against $H_1 : \mu \neq 5$ at 5%

$$test = \frac{(\bar{x} - 5)}{s/\sqrt{n}} = \frac{(4.1777 - 5)}{(3.3103)/\sqrt{428}} = -5.1392 \sim t_{427}$$

Since $|-5.1392| > 1.9655$ (the 97.5% quantile of the t_{427}), we reject H_0 . (1 point for the statistic and 1 point for the conclusion)

By default, we assume normality of the data, a size of 5% and a two-sided alternative. When we assume normality, the distribution used for the critical value is the t-distribution with $(n - 1)$ degrees of freedom. If we don't, the critical value is based on the asymptotic $N(0,1)$ property of the test. It is also possible to change the alternative hypothesis by either "greater" or "less" and the size of the test:

```
testMean(wage, h0=5, size=0.10, alter="less", assume="nonNormal")
```

Testing $H_0 : \mu = 5$ against $H_1 : \mu < 5$ at 10%

$$test = \frac{(\bar{x} - 5)}{s/\sqrt{n}} = \frac{(4.1777 - 5)}{(3.3103)/\sqrt{428}} = -5.1392 \approx N(0, 1)$$

Since $-5.1392 < -1.2816$ (the 10% quantile of the $N(0, 1)$), we reject H_0 . (The $N(0, 1)$ is an approximation based on the C.L.T. because the distribution of the data is unknown)

By default, all decimals are kept to compute the statistic. This could lead to slightly different solutions when the question is asked in an exam, because the printed numbers are rounded. It is possible to round the sample mean and standard errors before computing the test, through the argument `dround`. For example, suppose an exam question was generated directly from the data and the following table was printed using `stargazer`:

```
library(stargazer)
stargazer(data.frame(wage), digits=3, header=FALSE, float=FALSE)
```

Statistic	N	Mean	St. Dev.	Min	Max
wage	428	4.178	3.310	0.128	25.000

We would generate the solution as follows:

```
testMean(wage, h0=5, size=0.10, alter="less", dround=3)
```

Testing $H_0 : \mu = 5$ against $H_1 : \mu < 5$ at 10%

$$test = \frac{(\bar{x} - 5)}{s/\sqrt{n}} = \frac{(4.178 - 5)}{(3.31)/\sqrt{428}} = -5.1377 \sim t_{427}$$

Since $-5.1377 < -1.2835$ (the 10% quantile of the t_{427}), we reject H_0 .

This option exists for all solution generator, so we won't discuss it further. It also possible to generate a solution without data. We just need to provide the sample mean, the standard error and the sample size:

```
testMean(h0=4, xbar=3.7, se=0.8, n=40)
```

Testing $H_0 : \mu = 4$ against $H_1 : \mu \neq 4$ at 5%

$$test = \frac{(\bar{x} - 4)}{s/\sqrt{n}} = \frac{(3.7 - 4)}{(0.8)/\sqrt{40}} = -2.3717 \sim t_{39}$$

Since $|-2.3717| > 2.0227$ (the 97.5% quantile of the t_{39}), we reject H_0 .

3.1.2 Confidence interval for the mean

The $(1 - \alpha) \times 100\%$ confidence interval is defined as:

$$[\bar{X} - q_{1-\alpha/2}\hat{\sigma}, \bar{X} + q_{1-\alpha/2}\hat{\sigma}],$$

where $q_{1-\alpha/2}$ is the $(1 - \alpha/2) \times 100\%$ quantile of the t_{n-1} when $X_i \sim N(\mu, \sigma^2)$, in which case the coverage is exact, and the $N(0, 1)$ when the distribution of X_i is unknown. For the latter case, the coverage is just an approximation based on the CLT. This rule is consistent with the rule for tests on the mean described in the previous section.

If we use the wage data from the previous section, the 95% confidence interval for the average wage, assuming normality, is:

```
ciMean(wage)
```

95% confidence interval for the mean

$$\begin{aligned} CI &= \left[\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}} \right] \\ &= \left[4.1777 - 1.9655 \frac{3.3103}{\sqrt{428}}, 4.1777 + 1.9655 \frac{3.3103}{\sqrt{428}} \right] \\ &= [3.8632, 4.4922] \end{aligned}$$

The t^* is the 97.5% quantile of the t-distribution with 427 degrees of freedom.

As for `testMean`, we can choose not to assume normality and change the `size`. It is also possible to round the mean and standard deviation to match the solution of written questions.

```
ciMean(wage, size=0.15, assume="nonNormal", dround=3)
```

85% confidence interval for the mean

$$\begin{aligned} CI &= \left[\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}} \right] \\ &= \left[4.178 - 1.4395 \frac{3.31}{\sqrt{428}}, 4.178 + 1.4395 \frac{3.31}{\sqrt{428}} \right] \\ &= [3.9477, 4.4083] \end{aligned}$$

The t^* is the 92.5% quantile of the $N(0, 1)$ (The $N(0, 1)$ is an approximation based on the C.L.T. because the distribution of the data is unknown).

Finally, we can specify the sample mean, standard deviation and sample size:

```
ciMean(xbar=3.2, se=0.9, n=32)
```

95% confidence interval for the mean

$$\begin{aligned} CI &= \left[\bar{x} - t^* \frac{s}{\sqrt{n}}, \bar{x} + t^* \frac{s}{\sqrt{n}} \right] \\ &= \left[3.2 - 2.0395 \frac{0.9}{\sqrt{32}}, 3.2 + 2.0395 \frac{0.9}{\sqrt{32}} \right] \\ &= [2.8755, 3.5245] \end{aligned}$$

The t^* is the 97.5% quantile of the t-distribution with 31 degrees of freedom.

3.1.3 Tests on the difference between two means

The following properties are assumed in the solution generator `testDiffMeans`. If we have two samples of sizes n_1 and n_2 for X_1 and X_2 , and want test the hypothesis $H_0 : \mu_1 - \mu_2 = c$, then:

- If $X_1 \sim N(\mu_1, \sigma^2)$ and $X_2 \sim N(\mu_2, \sigma^2)$, which implies that we assume equal variance, we have the following result under the null:

$$\frac{\bar{X}_1 - \bar{X}_2 - c}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

where

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2} [\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)]$$

and $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are the usual bias corrected estimator of the variance of X_1 and X_2 .

- For any other cases, which include non-normality and/or non-equal variances, the exact distribution of the test is unknown, so we use the approximated $N(0,1)$ instead of the t-distribution. If the variances are not equal, the above test is not valid and must be replaced by:

$$\frac{\bar{X}_1 - \bar{X}_2 - c}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \approx N(0, 1).$$

Note: The test presented here assumes that $Cov(X_1, X_2) = 0$. We do not cover the non-zero covariance case in my courses, so it is not yet implemented.

Suppose we want to test if the average wage if the same for workers living in a city and the ones not living in a city, we can proceed as follows. We first consider the normal case with equal variances.

```
wageCity <- subset(PSID1976, hours>0 & city=="yes")$wage
wageNoCity <- subset(PSID1976, hours>0 & city=="no")$wage
testDiffMeans(x1=wageCity, x2=wageNoCity, h0=0)
```

Testing $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$ at 5%

$$\begin{aligned} s &= \sqrt{\frac{1}{n_1 + n_2 - 2} [\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\ &= \sqrt{\frac{1}{426} [13.4736 \times (274 - 1) + 6.1053 \times (154 - 1)] \left(\frac{1}{274} + \frac{1}{154} \right)} \\ &= 0.3314 \end{aligned}$$

$$test = \frac{(\bar{x}_1 - \bar{x}_2)}{s} = \frac{(4.4735 - 3.6513)}{0.3314} = 2.4813 \sim t_{426}$$

Since $|2.4813| > 1.9655$ (the 97.5% quantile of the t_{426}), we reject H_0 .

For any other cases, the approximated $N(0,1)$ is used. Here is an example with other specifications:

```
testDiffMeans(x1=wageCity, x2=wageNoCity, h0=1, assumev="diff", size=0.10,
              alter="less")
```

Testing $H_0 : \mu_1 - \mu_2 = 1$ against $H_1 : \mu_1 - \mu_2 < 1$ at 10%

$$\begin{aligned} s &= \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{13.4736}{274} + \frac{6.1053}{154}} = 0.298 \\ test &= \frac{(\bar{x}_1 - \bar{x}_2) - 1}{s} = \frac{(4.4735 - 3.6513) - 1}{0.298} = -0.5963 \approx N(0, 1) \end{aligned}$$

Since $-0.5963 > -1.2816$ (the 90% quantile of the $N(0,1)$), we do not reject H_0 . (The $N(0,1)$ is an approximation based on the C.L.T. because the distribution of the data is unknown)

As for the other tests, we can input estimated means and standard errors instead of vectors. The arguments `xbar`, `se`, and `n` must be vectors of two:

```
testDiffMeans(h0=1, xbar=c(2.2, 3.3), se=c(3.4, 4.6), n=c(34, 76))
```

Testing $H_0 : \mu_1 - \mu_2 = 1$ against $H_1 : \mu_1 - \mu_2 \neq 1$ at 5%

$$\begin{aligned}s &= \sqrt{\frac{1}{n_1 + n_2 - 2} [\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\&= \sqrt{\frac{1}{108} [11.56 \times (34 - 1) + 21.16 \times (76 - 1)] \left(\frac{1}{34} + \frac{1}{76} \right)} \\&= 0.8809\end{aligned}$$

$$test = \frac{(\bar{x}_1 - \bar{x}_2) - 1}{s} = \frac{(2.2 - 3.3) - 1}{0.8809} = -2.3841 \sim t_{108}$$

Since $|-2.3841| > 1.9822$ (the 97.5% quantile of the t_{108}), we reject H_0 .

3.1.4 Confidence intervals for the difference between two means

The theory from the previous section also applies here: we use the t-distribution only if the data is normally distributed and the variances of X_1 and X_2 are the same. Also, the standard deviation of $\bar{X}_1 - \bar{X}_2$ is

$$s = \sqrt{\frac{1}{n_1 + n_2 - 2} [\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)},$$

if the variances are the same and the following if they are not:

$$s = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}.$$

The confidence interval for the difference in average wage between workers from a city the ones not from a city, assuming normality and equal variance, is

```
ciDiffMeans(wageCity, wageNoCity)
```

95% confidence interval for $(\mu_1 - \mu_2)$

$$\begin{aligned}s &= \sqrt{\frac{1}{n_1 + n_2 - 2} [\hat{\sigma}_1^2(n_1 - 1) + \hat{\sigma}_2^2(n_2 - 1)] \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \\&= \sqrt{\frac{1}{426} [13.4736 \times (274 - 1) + 6.1053 \times (154 - 1)] \left(\frac{1}{274} + \frac{1}{154} \right)} \\&= 0.3314\end{aligned}$$

$$\begin{aligned}CI &= [(\bar{X}_1 - \bar{X}_2) - t^* s, (\bar{X}_1 - \bar{X}_2) + t^* s] \\&= [(4.4735 - 3.6513) - 1.9655 \times 0.3314, (4.4735 - 3.6513) + 1.9655 \times 0.3314] \\&= [0.1709, 1.4737]\end{aligned}$$

The t^* is the 97.5% quantile of the t-distribution with 426 degrees of freedom.

If we relax the normality and/or the equal variance we obtain:

```
ciDiffMeans(wageCity, wageNoCity, assumev="diff", size=0.1)
```

90% confidence interval for $(\mu_1 - \mu_2)$

$$s = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} = \sqrt{\frac{13.4736}{274} + \frac{6.1053}{154}} = 0.298$$

$$\begin{aligned} CI &= [(\bar{X}_1 - \bar{X}_2) - t^*s, (\bar{X}_1 - \bar{X}_2) + t^*s] \\ &= [(4.4735 - 3.6513) - 1.6449 \times 0.298, (4.4735 - 3.6513) + 1.6449 \times 0.298] \\ &= [0.3321, 1.3125] \end{aligned}$$

The t^* is the 95% quantile of the $N(0, 1)$ (The $N(0, 1)$ is an approximation based on the C.L.T. because the distribution of statistic is unknown).

As for testing the difference between means, we can also replace the vectors of observations by `xbar`, `se`, and `n`.

3.1.5 Test on the variance

The only implemented test at the moment is the one that assumes normality. The assumption implies that under the null $H_0 : \sigma^2 = c$, we have the following distribution:

$$\frac{(n-1)\hat{\sigma}^2}{c} \sim \chi_{n-1}^2$$

Let's consider the following summary statistics from the `hprice1` dataset:

```
data(hprice1, package="wooldridge")
hprice1$lotsize <- hprice1$lotsize/1000
hprice1$sqrft <- hprice1$sqrft/1000
stargazer(hprice1[,1:5], digits=4, header=FALSE, float=FALSE)
```

Statistic	N	Mean	St. Dev.	Min	Max
price	88	293.5460	102.7134	111.0000	725.0000
assess	88	315.7364	95.3144	198.7000	708.6000
bdrms	88	3.5682	0.8414	2	7
lotsize	88	9.0199	10.1742	1.0000	92.6810
sqrft	88	2.0137	0.5772	1.1710	3.8800

Suppose we want to test $H_0 : \sigma^2 = 130$ against $H_1 : \sigma^2 \neq 130$ for the lot size. In the following I show what happens if we set `assume` to "nonNormal". The test is performed, but a note is added saying that the chi-square is not a valid distribution.

```
testVar(hprice1$lotsize, h0=130, assume="nonNormal")
```

Testing $H_0 : \sigma^2 = 130$ against $H_1 : \sigma^2 \neq 130$ at 5%

$$test = \frac{(n-1)\hat{\sigma}^2}{130} = \frac{(87)103.5133}{130} = 69.2743 \sim \chi_{87}^2$$

Since $69.2743 > 63.0894$ (the 2.5% quantile of the χ_{87}^2) and $69.2743 < 114.6929$ (the 97.5% quantile of the χ_{87}^2), we do not reject H_0 . (The χ_{87}^2 is not valid in this case because the distribution of the data is unknown)

We can use one-sided tests and change the size:

```
testVar(hprice1$lotsize, h0=130, alter="less", size=0.1)
```

Testing $H_0 : \sigma^2 = 130$ against $H_1 : \sigma^2 < 130$ at 10%

$$test = \frac{(n - 1)\hat{\sigma}^2}{130} = \frac{(87)103.5133}{130} = 69.2743 \sim \chi_{87}^2$$

Since $69.2743 < 70.581$ (the 10% quantile of the χ_{87}^2), we reject H_0 .

```
testVar(hprice1$lotsize, h0=70, alter="greater", size=0.01)
```

Testing $H_0 : \sigma^2 = 70$ against $H_1 : \sigma^2 > 70$ at 1%

$$test = \frac{(n - 1)\hat{\sigma}^2}{70} = \frac{(87)103.5133}{70} = 128.6523 \sim \chi_{87}^2$$

Since $128.6523 > 120.591$ (the 99% quantile of the χ_{87}^2), we reject H_0 .

We can also replace the vector `par` values of `se` and `n`. For example, we want to use the number from the table and test $H_0 : \sigma^2 = 0.25$ for square footage, we proceed as follows:

```
testVar(se=0.5772, n=88, h0=0.25)
```

Testing $H_0 : \sigma^2 = 0.25$ against $H_1 : \sigma^2 \neq 0.25$ at 5%

$$test = \frac{(n - 1)\hat{\sigma}^2}{0.25} = \frac{(87)0.3332}{0.25} = 115.9396 \sim \chi_{87}^2$$

Since $115.9396 > 114.6929$ (the 97.5% quantile of the χ_{87}^2), we reject H_0 .