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)
```

$$\begin{aligned}\widehat{wage} = \underset{(3.2900)}{-8.5590} &+ \underset{(0.0495)}{0.4538}\ education + \underset{(0.1524)}{0.2558}\ age - \underset{(0.0018)}{0.0029}\ I(age^2)\\ n = 753, &\quad R^2 = 0.1047, SSR = 7075.373\end{aligned}$$

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, pchausse@uwaterloo.ca

$$
\begin{aligned}
\widehat{wage} = & \underset{(3.34990)^{**}}{-6.83341} + \underset{(0.09077)^{***}}{0.45828}\ education - \underset{(1.27155)}{1.42391}\ cityyes - \underset{(0.04801)^{**}}{0.12139}\ heducation \\
& + \underset{(0.15272)}{0.24075}\ age - \underset{(0.00176)}{0.00284}\ I(age^2) + \underset{(0.10318)}{0.13427}\ education : cityyes \\
& n = 753, \quad R^2 = 0.11471, SSR = 6996.553, \bar{R}^2 = 0.10759 \\
& {}^{*}\text{pv} < 0.1;\ {}^{**}\text{pv} < 0.05;\ {}^{***}\text{pv} < 0.01
\end{aligned}
$$

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

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

$$
\begin{aligned}
\widehat{wage} = & \underset{(3.2518)^{**}}{-6.8334} + \underset{(0.0719)^{***}}{0.4583}\ education - \underset{(1.1819)}{1.4239}\ cityyes - \underset{(0.0419)^{***}}{0.1214}\ heducation \\
& + \underset{(0.1559)}{0.2408}\ age - \underset{(0.0018)}{0.0028}\ I(age^2) + \underset{(0.1002)}{0.1343}\ education : cityyes \\
& n = 753, \quad R^2 = 0.1147, SSR = 6996.553 \text{ (Robust S-E)} \\
& {}^{*}\text{pv} < 0.1;\ {}^{**}\text{pv} < 0.05;\ {}^{***}\text{pv} < 0.01
\end{aligned}
$$

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:

```r
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)
```

$$
\begin{aligned}
\text{link}\left[\widehat{children}\right] = & \underset{(0.2557)^{***}}{-3.8200} - \underset{(0.0042)^{***}}{0.0187}\ educ + \underset{(0.0148)^{***}}{0.2650}\ age - \underset{(0.0002)^{***}}{0.0031}\ I(age^2) \\
& + \underset{(0.0414)}{0.0104}\ catholic - \underset{(0.0456)^{*}}{0.0803}\ electric + \underset{(0.0279)}{0.0444}\ radio \\
& - \underset{(0.0548)^{**}}{0.1229}\ tv - \underset{(0.0036)^{***}}{0.0126}\ heduc \\
& n = 1953, \quad AIC = 7344.376, \quad \text{Residual Deviance} = 1834.816, \\
& \text{Null Deviance} = 3246.903, \quad \text{Family} = \text{poisson}, \quad \text{Link} = \log \\
& {}^{*}\text{pv} < 0.1;\ {}^{**}\text{pv} < 0.05;\ {}^{***}\text{pv} < 0.01
\end{aligned}
$$

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:

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

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

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"`.

The purpose of these functions is to generate questions for assignments, exercises or exams and have a fast way to produce a detailed solution document. The questions can be generated by datasets, in which case the solutions are derived from the same dataset, or by made up numbers.

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:

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

were $q_{1-\alpha/2}$ 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 inn 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

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

  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%

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

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

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%

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

$$test = \frac{(\bar{x}_1 - \bar{x}_2) - 1}{s} = \frac{(4.4735 - 3.6513) - 1}{0.298} = -0.5963 \approx N(0,1)$$

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%

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

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

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

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

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 $(\mu_1 - \mu_2)$

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

$$
\begin{aligned}
CI &= \left[(\bar{X}_1-\bar{X}_2)-t^*s,\ (\bar{X}_1-\bar{X}_2)+t^*s\right] \\
&= \left[(4.4735-3.6513)-1.9655\times0.3314,\ (4.4735-3.6513)+1.9655\times0.3314\right] \\
&= [0.1709,\ 1.4737]
\end{aligned}
$$

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:

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

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

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

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

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:

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

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

```r
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.

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

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

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

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

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

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

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

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

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

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

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

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

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

We can also replace the numeric vector by 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:

```r
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%

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

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

3.1.6 Confidence intervals for the variance

As for testing the variance in the previous section, the confidence intervals presented here are only valid under normality. It is based on the same chi-square distribution. For example, if we want to generate a confidence interval for the variance of `sqrft` using the data from the previous table, we can do it using the dataset (setting `dround` to 4 as in the table):

```r
ciVar(x=hprice1$sqrft, dround=4)
```

95% confidence interval for the variance

$$\begin{aligned} CI &= \left[\frac{(n-1)S^2}{Q_u^*}, \frac{(n-1)S^2}{Q_l^*}\right] \\ &= \left[87 \times \frac{0.3332}{114.6929}, \ 87 \times \frac{0.3332}{63.0894}\right] \\ &= [0.2527, \ 0.4595] \end{aligned}$$

The Q_u^* is the 97.5% quantile of the χ^2_{87} and Q_l^* is its 2.5% quantile.

This is equivalent to using the numbers from the table:

```r
ciVar(se=0.5772, n=88)
```

95% confidence interval for the variance

$$
\begin{aligned}
CI &= \left[\frac{(n-1)S^2}{Q_u^*}, \frac{(n-1)S^2}{Q_l^*}\right] \\
&= \left[87 \times \frac{0.3332}{114.6929},\ 87 \times \frac{0.3332}{63.0894}\right] \\
&= [0.2527,\ 0.4594]
\end{aligned}
$$

The Q_u^* is the 97.5% quantile of the χ^2_{87} and Q_l^* is its 2.5% quantile.

As for the test, we can change the `size` for a different confidence level, and set `assume` to `"nonNormal"` to add a note about the distribution not being valid under non-normality.

3.1.7 Testing equality between two variances

We want to test $H_0:\ \sigma_1^2 = \sigma_2^2$ against $H_1:\ \sigma_1^2 \neq \sigma_2^2$ at $\alpha \times 100\%$. The only test I cover in the intro to statistics is the one that assumes normality of the two samples and their independence. Under these assumptions and the null hypothesis, we have the following distribution:

$$
\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F(n_1 - 1, n_2 - 1)
$$

For this test to be two-sided, we define $\hat{\sigma}_1^2$ as the highest estimate of the variance between the two samples. We reject of the statistic exceeds the $(1-\alpha)\times 100\%$ quantile of the distribution.

Consider the following summary statistics. The first table is for female workers and the second is not male workers. The dataset used to construct those tables is `wage1` form the `wooldridge` package.

Statistic	N	Mean	St. Dev.	Min	Max
wage	252	4.588	2.529	0.530	21.630
educ	252	12.317	2.473	0	18
exper	252	16.429	13.653	1	50
tenure	252	3.615	5.358	0	34

Statistic	N	Mean	St. Dev.	Min	Max
wage	274	7.099	4.161	1.500	24.980
educ	274	12.788	3.003	2	18
exper	274	17.558	13.500	1	51
tenure	274	6.474	8.369	0	44

If we assume that wage is normally distributed and that female wage is independent of male wage, we can test if the variance of wage is equal between the two groups. First, using the dataset `wage1` with `dround=3` to match the precision of the table:

```r
wF <- subset(wage1, female==1)$wage
wM <- subset(wage1, female==0)$wage
testDiffVar(wF, wM, dround=3)
```

Testing $H_0:\ \sigma_1^2 = \sigma_2^2$, against $H_1:\ \sigma_1^2 \neq \sigma_2^2$ at 5%

$$
test = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{17.3139}{6.3958} = 2.7071 \sim F(273, 251)
$$

Since $2.7071 > 1.2267$ (the 95% quantile of the $F(273, 251)$), we reject H_0.

We can also enter the information manually and change the size of the test:

```
testDiffVar(se=c(2.529, 4.161), n=c(252, 274), size=0.1)
```

Testing $H_0: \ \sigma_1^2 = \sigma_2^2$, against $H_1: \ \sigma_1^2 \neq \sigma_2^2$ at 10%

$$test = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{17.3139}{6.3958} = 2.7071 \sim F(273, 251)$$

Since $2.7071 > 1.1725$ (the 90% quantile of the $F(273, 251)$), we reject H_0.

3.2 Introduction to Econometrics

In this section, we present solution generators for inference based on linear regressions. The general model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

By default, the estimated covariance matrix used to perform tests and construct confidence intervals is the one obtained by `vcov`. When a robust estimate is needed, it has to be obtained outside the solution generator. As for inference on the means, the normality assumption implies exact tests using either the t or the F distribution. Without this assumption, the asymptotic standard normal or chi-square distribution is used.

3.2.1 Tests on a single regression coefficient

Consider the following estimated model:

```
fit <- lm(wage~educ+exper, wage1)
printReg(fit)
```

$$\widehat{wage} = \underset{(0.7666)}{-3.3905} \ + \ \underset{(0.0538)}{0.6443} \ educ \ + \ \underset{(0.0110)}{0.0701} \ exper$$
$$n = 526, \quad R^2 = 0.2252, SSR = 5548.16$$

If we want to test the hypothesis $H_0: \ \beta_1 = 0.5$ against the alternative $H_1: \ \beta_1 \neq 0.5$ with the assumption that the errors are conditionally normal and homoskedastic, we can run the following. Note that β_1 is the second coefficient, so we set `which` to 2. Also, to generate a solution used the rounded coefficient and standard error shown in the above regression, we set `dround` to 4:

```
testRegCoef(fit, which=2, h0=0.5, dround=4)
```

Testing $H_0: \ \beta_1 = 0.5$ against $H_1: \ \beta_1 \neq 0.5$ at 5%

$$test = \frac{(\beta_1 - 0.5)}{se} = \frac{(0.6443 - 0.5)}{0.0538} = 2.6822 \sim t_{523}$$

Since $|2.6822| > 1.9645$ (the 97.5% quantile of the t_{523}), we reject H_0.

We can also obtain the solution without the estimated model, by setting the coefficient estimate, its standard error, and sample size and the number of coefficients to the values we see in the printed equation:

```
testRegCoef(h0=0.5, beta=0.6443, se=0.0538, n=526, ncoef=3)
```

Testing $H_0: \ \beta_j = 0.5$ against $H_1: \ \beta_j \neq 0.5$ at 5%

$$test = \frac{(\beta_j - 0.5)}{se} = \frac{(0.6443 - 0.5)}{0.0538} = 2.6822 \approx N(0, 1)$$

Since $|2.6822| > 1.96$ (the 97.5% quantile of the $N(0,1)$), we reject H_0. (The N(0,1) is an approximation based on the C.L.T. because either the distribution of the error term is unknown and/or the errors are not homoskedastic.)

If we relax the normality assumption, the critical value is obtained using the N(0,1). If a vector of standard errors is provided, the solution generator assumes that the ones obtained with `vcov` is considered non-valid. In this case, there are no exact distribution, so the N(0,1) is also used. In the following the test is performed using the HCCM standard errors. For a solution based on a printed regression, we have to print the regression with the robust standard errors.

```
rse <- sqrt(diag(vcovHC(fit)))
printReg(fit, se=rse)
```

$$\begin{aligned}\widehat{wage} = \underset{(0.8748)}{-3.3905} \; + \; \underset{(0.0659)}{0.6443}\; educ \; + \; \underset{(0.0111)}{0.0701}\; exper \\ n = 526, \quad R^2 = 0.2252, SSR = 5548.16 \text{ (Robust S-E)}\end{aligned}$$

In the following test, the N(0,1) is used instead of the t distribution, because it is assume that the homoskedasticity assumption does not hold.

```
testRegCoef(fit, which=2, h0=0.5, dround=4, se=rse)
```

Testing $H_0: \ \beta_1 = 0.5$ against $H_1: \ \beta_1 \neq 0.5$ at 5%

$$test = \frac{(\beta_1 - 0.5)}{se} = \frac{(0.6443 - 0.5)}{0.0659} = 2.1897 \approx N(0,1)$$

Since $|2.1897| > 1.96$ (the 97.5% quantile of the $N(0,1)$), we reject H_0. (The N(0,1) is an approximation based on the C.L.T. because either the distribution of the error term is unknown and/or the errors are not homoskedastic.)

As for tests on the mean, we can change the size with the argument `size` and the alternative with the argument `alter`.

3.2.2 Confidence interval on a regression coefficient

The same rule applies for confidence intervals. It is exact (using the t distribution) only if we assume normality and do not provide alternative standard errors. For example, we can obtain the 90% confidence interval for the coefficient of `bdrms` in the following regression as follows. First the estimated model is printed:

```
fit <- lm(log(price)~bdrms+log(lotsize)+log(sqrft), hprice1)
printReg(fit)
```

$$\begin{aligned}\widehat{log(price)} = \underset{(0.1049)}{4.7003} \; + \; \underset{(0.0275)}{0.0370}\; bdrms \; + \; \underset{(0.0383)}{0.1680}\; log(lotsize) \; + \; \underset{(0.0929)}{0.7002}\; log(sqrft) \\ n = 88, \quad R^2 = 0.643, SSR = 2.8626\end{aligned}$$

Then we construct the interval:

```
ciRegCoef(fit, which=2, size=0.10, dround=4)
```

90% confidence interval for β_1.

$$\begin{aligned} CI &= [\beta_1 - t^* se, \ \beta_1 + t^* se] \\ &= [0.037 - 1.6632 \times 0.0275, \ 0.037 + 1.6632 \times 0.0275] \\ &= [-0.0087, \ 0.0827] \end{aligned}$$

The t^* is the 95% quantile of the t-distribution with 84 degrees of freedom.

To get confidence interval with robust standard errors or non-normality, we just change `se` and `assume` as for the solution of hypothesis tests.

3.2.3 Testing linear combinations of coefficients (direct-t)

Any test on linear combinations of coefficients can be written as $H_0: R\beta = q$. In this section we consider single hypotheses, so R has one row and q is a scalar. For the function `testLinRegCoefs`, R can be a plain vector or a matrix. The only requirement is to have a length equal to the number of coefficients. Consider the following regression, followed by the estimated covariance matrix:

```
dat <- subset(wage1, female==1 & married==1 & smsa==1)

fit <- lm(wage~educ+exper+tenure, dat)
printReg(fit, digits=5)
```

$$\widehat{wage} = \underset{(1.37570)}{1.73810} + \underset{(0.10038)}{0.24617}\, educ - \underset{(0.01712)}{0.00470}\, exper - \underset{(0.03666)}{0.00235}\, tenure$$
$$n = 88, \quad R^2 = 0.07187, SSR = 273.8005$$

```
knitr::kable(round(vcov(fit),5))
```

	(Intercept)	educ	exper	tenure
(Intercept)	1.89256	-0.13345	-0.00777	0.00117
educ	-0.13345	0.01008	0.00026	-0.00015
exper	-0.00777	0.00026	0.00029	-0.00027
tenure	0.00117	-0.00015	-0.00027	0.00134

If we want to generate a reproducible solution, we need to set `dround` to 5, which is the number of digits shown in the regression and covariance matrix. In the following, we test the hypothesis $\beta_1 - 10\beta_2 = 0.6$, which implies $R = \{0, 1, -10, 0\}$ and $q = 0.6$.

```
R <- c(0,1,-10,0)
q <- 0.6
testLinRegCoefs(object=fit, R=R, q=q, dround=5, alter="less")
```

Testing $H_0: \ \beta_1 - 10\beta_2 = 0.6$ against $H_1: \ \beta_1 - 10\beta_2 < 0.6$ at 5%

$$test = \frac{\hat{\beta}_1 - 10\hat{\beta}_2 - 0.6}{se} = \frac{-0.3068}{0.1841} = -1.667 \sim t_{84}$$

where,

$$se = \sqrt{Var(\hat{\beta}_1) + 10^2 Var(\hat{\beta}_2) - 2 \times 10 Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

Since $-1.667 < -1.6632$ (the 5% quantile of the t_{84}), we reject H_0.

If we set `assume` to `"nonNormal"` , or input an alternative covariance matrix, the N(0,1) will be used to compute the critical values.

```
testLinRegCoefs(object=fit, R=R, q=q, dround=5, alter="greater",
                vcoef=vcovHC(fit))
```

Testing $H_0: \ \beta_1 - 10\beta_2 = 0.6$ against $H_1: \ \beta_1 - 10\beta_2 > 0.6$ at 5%

$$test = \frac{\hat{\beta}_1 - 10\hat{\beta}_2 - 0.6}{se} = \frac{-0.3068}{0.1667} = -1.8406 \approx N(0,1)$$

where,

$$se = \sqrt{Var(\hat{\beta}_1) + 10^2 Var(\hat{\beta}_2) - 2 \times 10 Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

Since $-1.8406 < 1.6449$ (the 95% 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 either the distribution of the error term is unknown and/or the errors are not homoskedastic.)

It is also possible to generate a solution with specific covariance matrix and coefficient vector. In this case, we need to specify if we assume normality or not, in order to base the test on the t distribution or the N(0,1). The following is the same solution without the `lm` object, but with the t-distribution:

```
v <- round(vcovHC(fit),5)
b <- round(coef(fit),5)
testLinRegCoefs(R=R, q=q, alter="greater", vcoef=v,  beta=b, n=nobs(fit))
```

Testing $H_0: \ \beta_1 - 10\beta_2 = 0.6$ against $H_1: \ \beta_1 - 10\beta_2 > 0.6$ at 5%

$$test = \frac{\hat{\beta}_1 - 10\hat{\beta}_2 - 0.6}{se} = \frac{-0.3068}{0.1667} = -1.8406 \approx N(0,1)$$

where,

$$se = \sqrt{Var(\hat{\beta}_1) + 10^2 Var(\hat{\beta}_2) - 2 \times 10 Cov(\hat{\beta}_1, \hat{\beta}_2)}$$

Since $-1.8406 < 1.6449$ (the 95% 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 either the distribution of the error term is unknown and/or the errors are not homoskedastic.)

3.2.4 Testing multiple restrictions with the F test

The solution generator `testRegF` is for testing joint hypotheses under the homoskedasticity assumption. I don't ask questions about the non-homoskedastic case in exams, because it involves matrix calculations that are not suite for questions to be answered in short time. That's why this case is not covered. The test under normal errors is:

$$\frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F(q, n-k-1)\,,$$

where q is the number of restrictions being tested, SSR_u is the unrestricted SSR and SSR_r is the restricted SSR. If we drop the normality assumption, we use the following approximation based on the C.L.T.:

$$\frac{(SSR_r - SSR_u)}{SSR_u/(n-k-1)} \approx \chi^2_q$$

We do not present solutions based on the restricted and unrestricted R^2, because its validity depends on left-hand side of the regression being the same in the restricted and unrestricted models. However, the solution of the joint significance of all coefficients but the intercept is based on the R^2 of the unrestricted model. We have the following with and without the normality assumption:

$$\frac{R^2/(k-1)}{(1-R^2)/(n-k-1)} \sim F(k-1, n-k-1)$$

$$\frac{R^2}{(1-R^2)/(n-k-1)} \approx \chi^2_{k-1}$$

Consider the following models:

```r
data(hprice3, package="wooldridge")
hprice3 <- subset(hprice3, year==1978)
hprice3$price <- hprice3$price/1000
fitu <- lm(price~age+I(age^2)+rooms*baths+area+land, hprice3)
fitr <- lm(price~age+rooms+baths+area+land, hprice3)
printReg(fitu, maxpl=3, digits=5)
```

$$\begin{aligned}\widehat{price} = &\underset{(27.53351)}{39.94059} - \underset{(0.13627)}{0.56601}\ age + \underset{(0.00084)}{0.00236}\ I(age^2) - \underset{(4.28837)}{2.34911}\ rooms \\ &- \underset{(12.78385)}{14.43260}\ baths + \underset{(0.00323)}{0.02322}\ area + \underset{(0.00003)}{0.00011}\ land \\ &+ \underset{(1.89571)}{2.68518}\ rooms : baths \\ n = 179,& \quad R^2 = 0.62298, SSR = 62948.16\end{aligned}$$

```r
printReg(fitr, maxpl=3, digits=5)
```

$$\begin{aligned}\widehat{price} = &\underset{(10.42160)}{-13.58579} - \underset{(0.04365)}{0.18786}\ age + \underset{(1.99856)}{4.15752}\ rooms + \underset{(3.18275)}{6.15880}\ baths \\ &+ \underset{(0.00329)}{0.02374}\ area + \underset{(0.00003)}{0.00012}\ land \\ n = 179,& \quad R^2 = 0.59966, SSR = 66840.94\end{aligned}$$

We want to test the significance of all coefficients but the intercept under normality. We can either provide the `lm` object of the unrestricted model, or give the values of the R^2, sample size and the number of coefficients:

```r
testRegF(fitu, dround=5)
```

Testing joint hypotheses using the F-test

$$test = \frac{R^2/7}{(1-R^2)/171} = \frac{0.623/7}{(1-0.623)/171} = 40.3653 \sim F(7, 171)$$

Since $40.3653 > 2.0635$ (the 95% quantile of the $F(7, 171)$), we reject H_0.

```r
testRegF(Rsq=0.62298, n=179, ncoef=8)
```

Testing joint hypotheses using the F-test

$$test = \frac{R^2/7}{(1-R^2)/171} = \frac{0.623/7}{(1-0.623)/171} = 40.3653 \sim F(7, 171)$$

Since $40.3653 > 2.0635$ (the 95% quantile of the $F(7, 171)$), we reject H_0.

If we want to test the second model against the first, which corresponds to $H_0: \ \beta_2 = \beta_7 = 0$, we proceed as follows (`dround` is set to 2, because only two decimals of the SSR's are printed). To see what happens if we do not assume normality, we set `assume` to `"nonNormal"`.

```r
testRegF(fitu, fitr, dround=2, assume="nonNormal")
```

Testing joint hypotheses using the F-test

$$test = \frac{(SSR_r - SSR_{ur})}{SSR_{ur}/171} = \frac{(66840.94 - 62948.16)}{62948.16/171} = 10.5748 \approx \chi^2_2$$

Since $10.5748 > 5.9915$ (the 95% quantile of the χ^2_2), we reject H_0. (The χ^2_2 is an approximation based on the C.L.T. because the distribution of the error term is unknown.

It is also possible to just provide the values of the SSR's and change the size:

```
testRegF(assume="Normal", size=0.1, SSRu=62948.16, SSRr=66840.94,
         n=179, nrest=2, ncoef=8)
```

Testing joint hypotheses using the F-test

$$test = \frac{(SSR_r - SSR_{ur})/2}{SSR_{ur}/171} = \frac{(66840.94 - 62948.16)/2}{62948.16/171} = 5.2874 \sim F(2, 171)$$

Since $5.2874 > 2.3339$ (the 90% quantile of the $F(2, 171)$), we reject H_0.

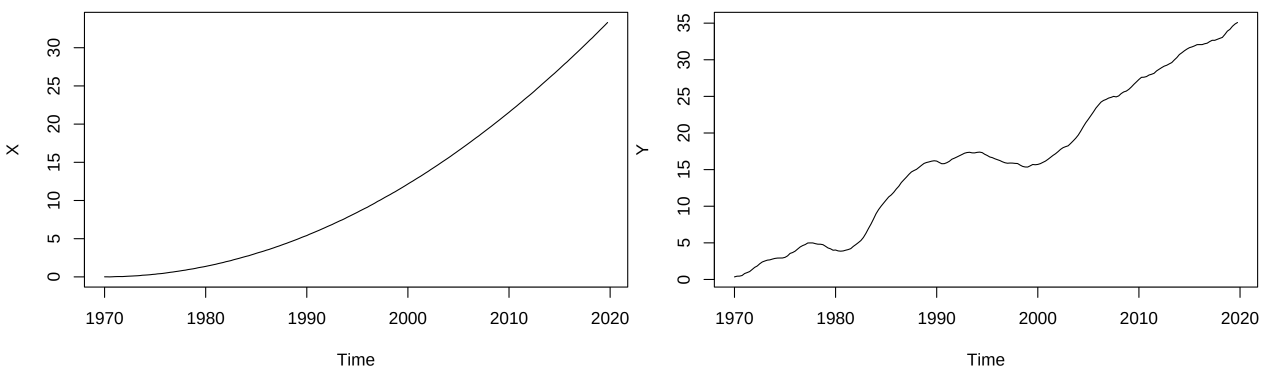
4 Miscellaneous

In this section, I present some functions I used in the course as tools for teaching or to generate random assignments.

4.1 Generate time series

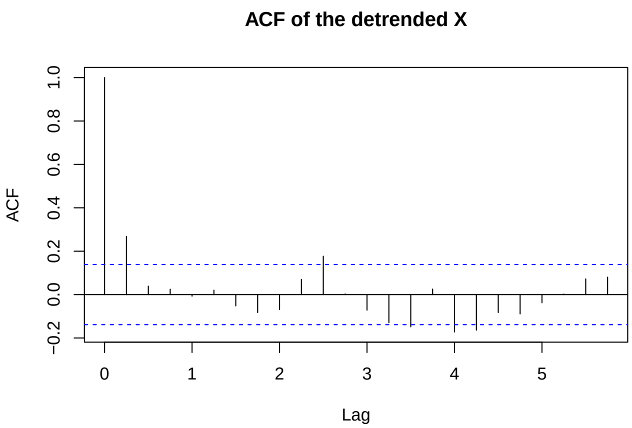
The function `genTS` generates an `mts` object with two time series, `X` and `Y`. One of the series is trend stationary (TS) and the second one is a unit root process with possibly a constant or a linear drift. The purpose of this dataset is to identify the process for each series. The following is an example with all possible arguments, which are all optional.

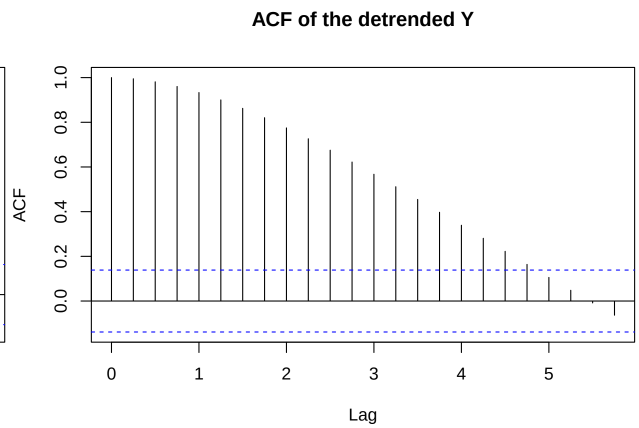
```
dat <- genTS(n=200, frequency=4, start=1970, seed=112345)
plot(dat[,"X"], ylab="X")
plot(dat[,"Y"], ylab="Y")
```



Since we know the possible processes, it is not too hard to determine the process of each series:

```
fitX <- lm(X~time(X) + I(time(X)^2), dat)
Xdt <- ts(residuals(fitX), start=start(dat), frequency=frequency(dat))
fitY <- lm(Y~time(Y) + I(time(Y)^2), dat)
Ydt <- ts(residuals(fitY), start=start(dat), frequency=frequency(dat))
acf(Xdt, main="ACF of the detrended X")
acf(Ydt, main="ACF of the detrended Y")
```





We would conclude that since the detrended Y is highly persistent, it is probably a unit root process, and X would be a trend stationary series. The true processes are saved in the attribute `"which"`:

```
attr(dat, "which")
```

```
## [1] "X is a  TS"          "Y is a  Unit Root"
```

4.2 Generate supply and demand functions

```
sys <- genDemSup(200)
sys
```

$$
\begin{aligned}
Q^d &= 100 - 1P + 2X + \eta \\
Q^s &= 20 + 1P - 2Z + e
\end{aligned}
$$

```
plot(sys, nCurves=15)
```