Goodness–of–fit tests in semi–linear models

Jochen Einbeck

Department of Mathematical Sciences, Durham University

jochen.einbeck@durham.ac.uk

joint work with Simos Meintanis (University of Athens)

London, 10th December 2010

Durham University

Semi–linear model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + g(\boldsymbol{z}) + \sigma\varepsilon$$

where

- $\boldsymbol{x} = (x_1, \ldots, x_p)'$ and $\boldsymbol{z} = (z_1, \ldots, z_q)'$ are multivariate covariates;
- $g : \mathbb{R}^q \longrightarrow \mathbb{R}$ is a smooth, unspecified function.
- the error ε follows an unknown distribution F, with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{E}(\varepsilon^2) = 1$;
- $\boldsymbol{\beta}$, $g(\cdot)$, and σ have to be estimated from iid observations

$$\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i\} \in \mathbb{R}^{1+p+q}, \quad i = 1, \ldots, n.$$

Semi–linear model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + g(\boldsymbol{z}) + \sigma\varepsilon$$

where

- $\boldsymbol{x} = (x_1, \ldots, x_p)'$ and $\boldsymbol{z} = (z_1, \ldots, z_q)'$ are multivariate covariates;
- $g : \mathbb{R}^q \longrightarrow \mathbb{R}$ is a smooth, unspecified function.
- the error ε follows an unknown distribution F, with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{E}(\varepsilon^2) = 1$;
- $\boldsymbol{\beta}$, $g(\cdot)$, and σ have to be estimated from iid observations

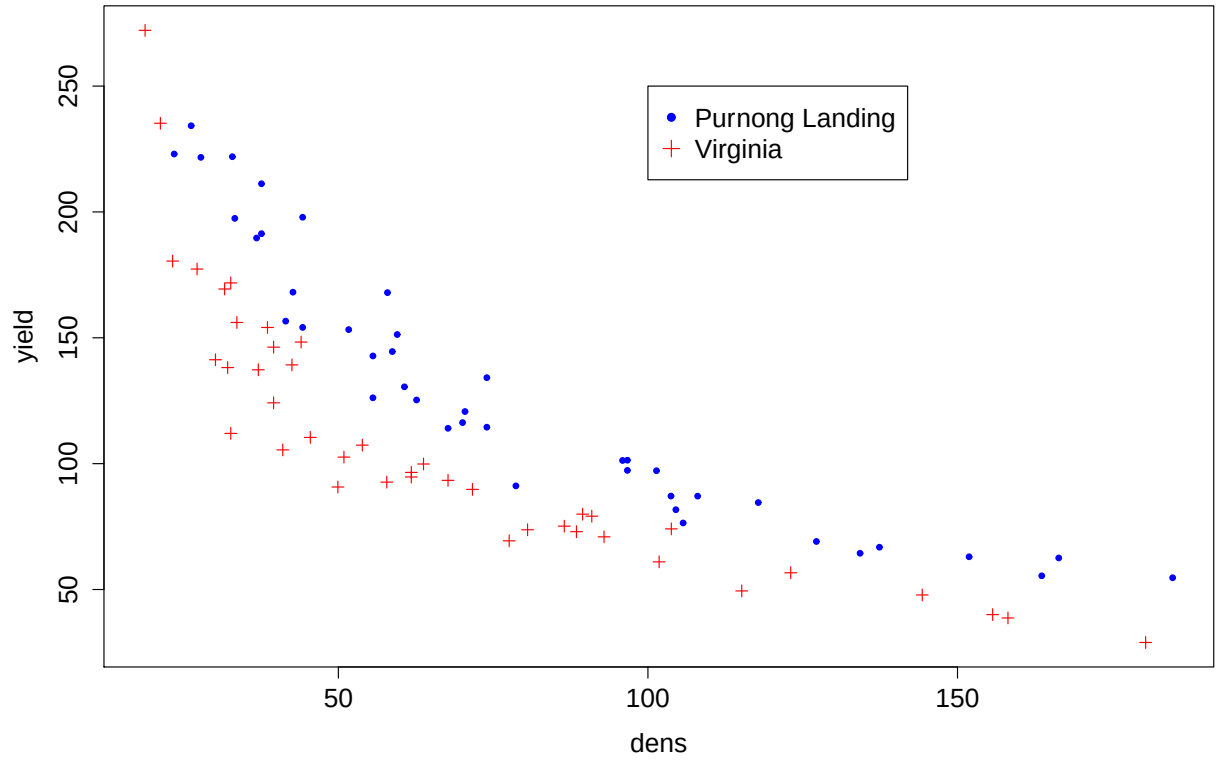
$$\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i\} \in \mathbb{R}^{1+p+q}, \quad i = 1, \ldots, n.$$

Important subcases:

- the linear model for $q = 0$;
- the nonparametric regression model for $p = 0$;
- the partial linear model for $p \geq 1$ and $q = 1$.

Example: Onions data (cont.)

- 84 observations from an experiment involving the production of white Spanish onions in two South Australian locations.
- Plotted is onion `yield` in grammes per plant vs. areal density of plants (plants per square metre):

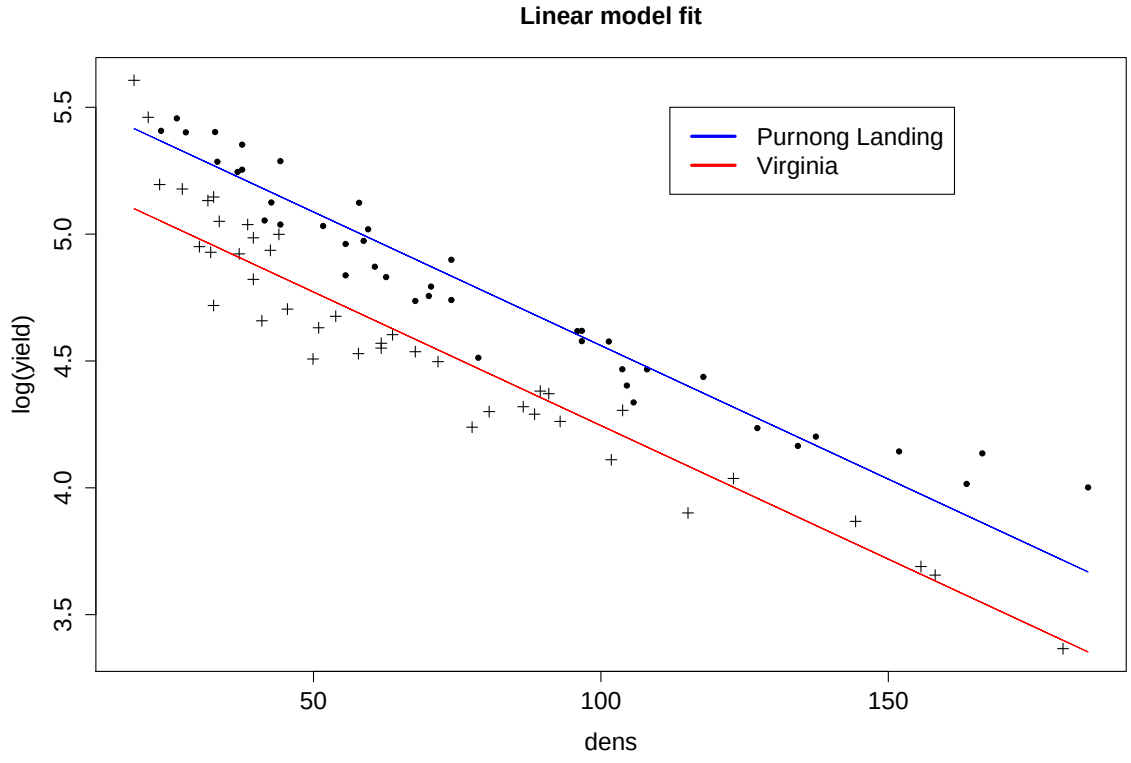


Example: Onions data (cont.)

- Linear model:

$$\log(\texttt{yield}_i) = \beta_0 + \beta_1 \texttt{location}_i + \beta_2 \texttt{dens}_i$$

where $\texttt{location}_i = 1_{\{i-\text{th obs. from Purnong Landing}\}}$.

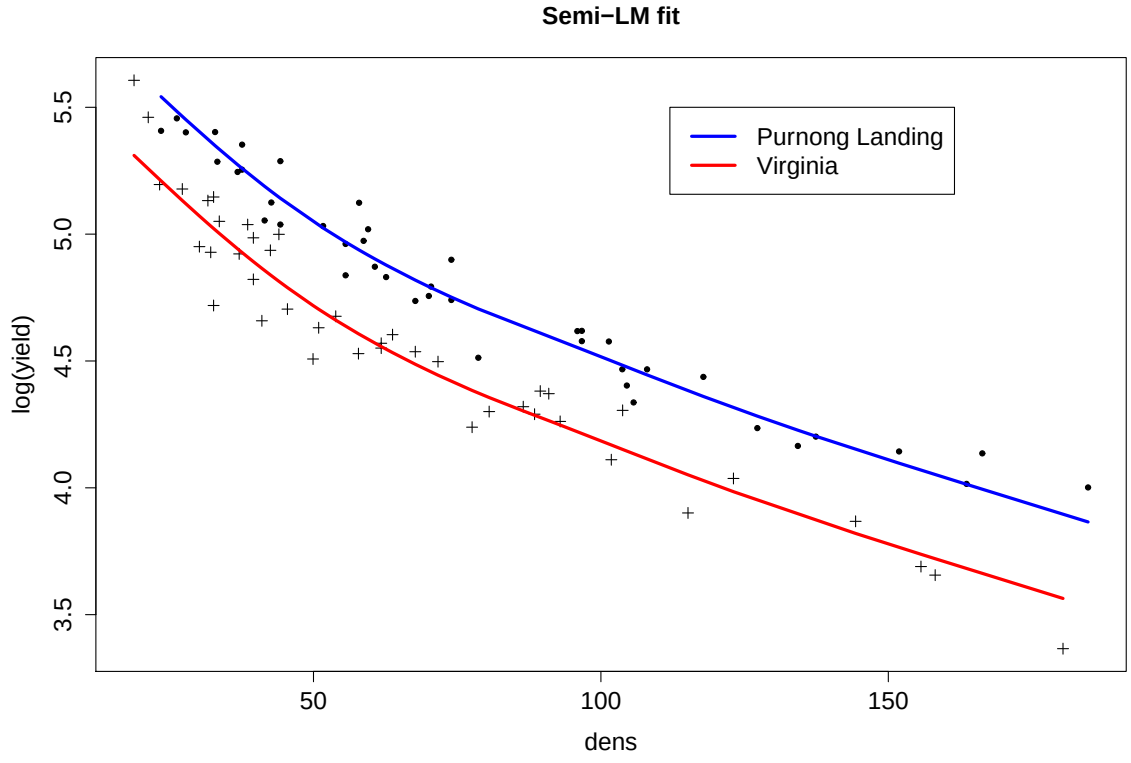


Example: Onions data (cont.)

- Semi-linear model:

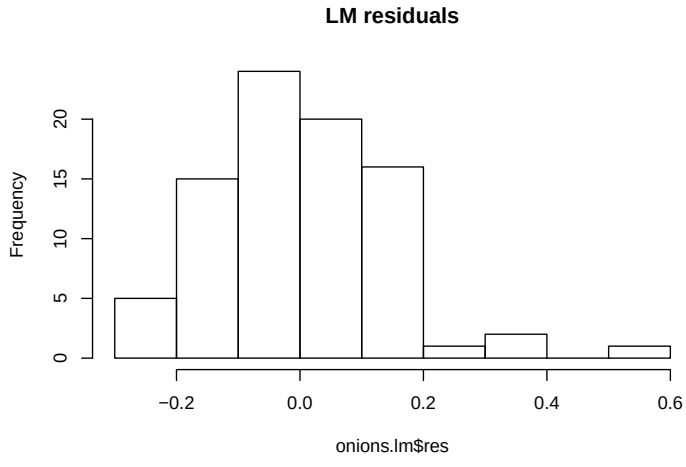
$$\log(\texttt{yield}_i) = \beta_0 + \beta_1 \texttt{location}_i + \beta_2 g(\texttt{dens}_i)$$

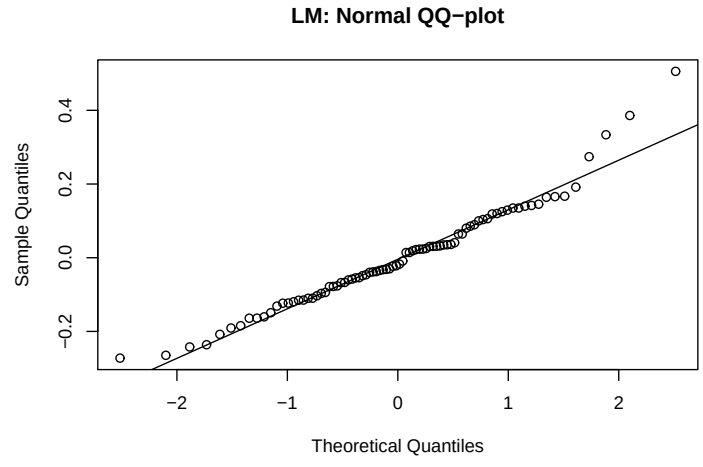
where $\texttt{location}_i = 1_{\{i-\text{th obs. from Purnong Landing}\}}$.

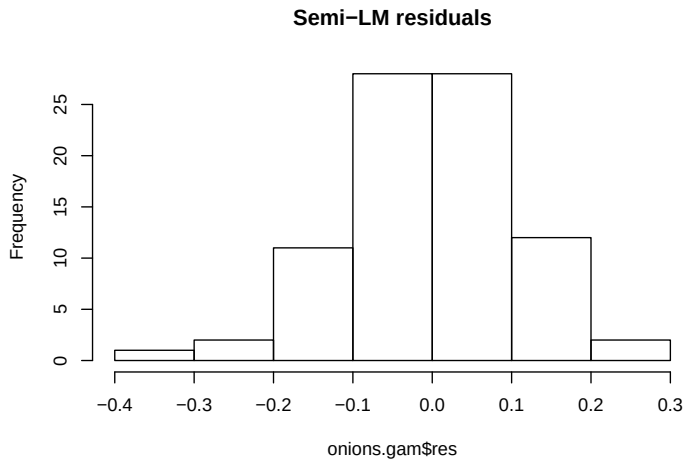


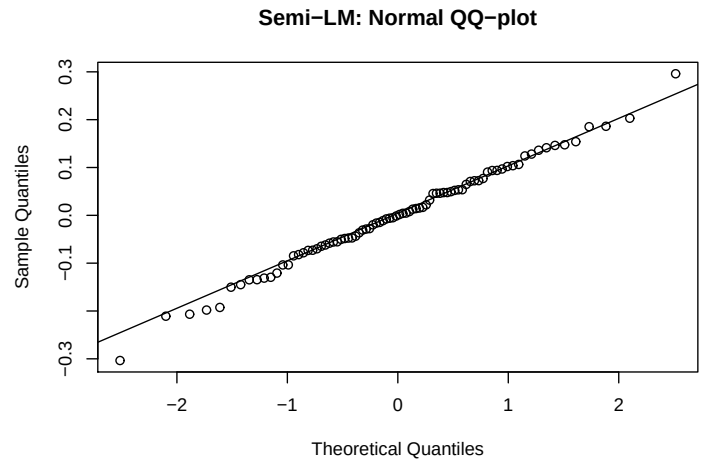
Example: Onions data (cont.)

- Consider residual distribution:



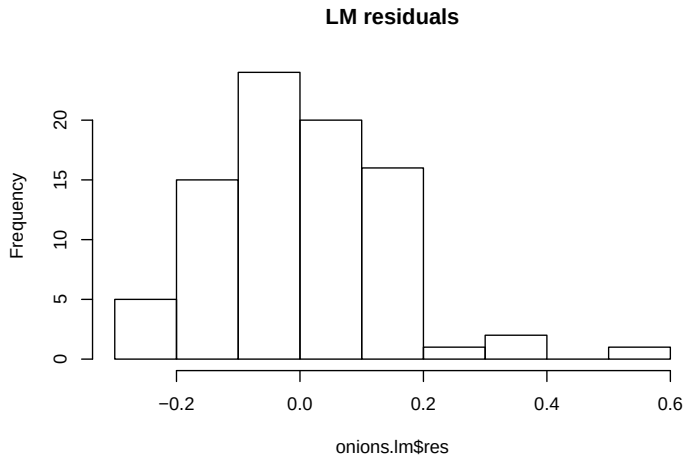


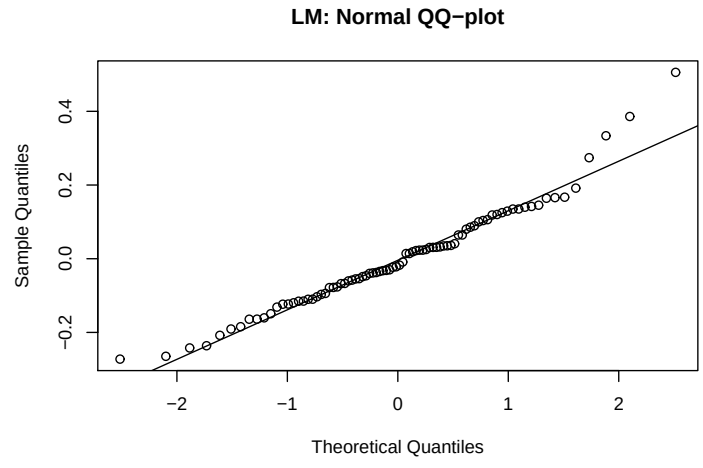


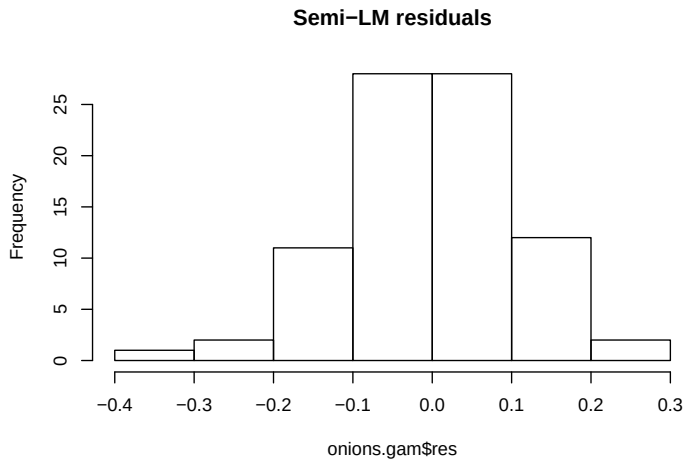


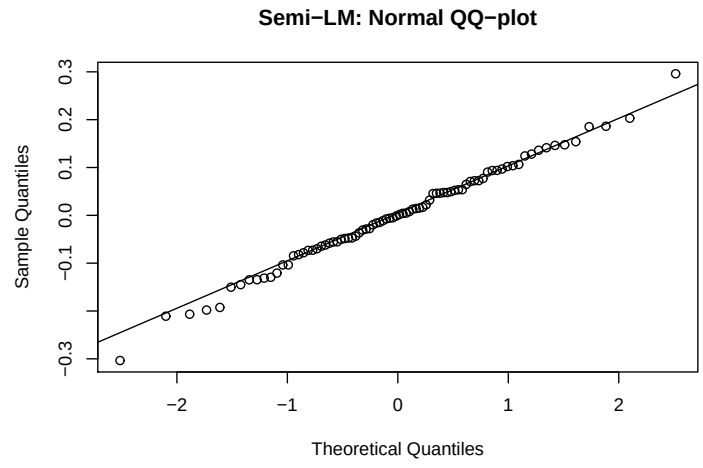
Example: Onions data (cont.)

- Consider residual distribution:









- The latter looks more 'normal', but can we test for this formally?

Tests for the error distribution

We wish to test for

- the specific parametric form of the error distribution F, i.e. whether or not it corresponds to a specific family of distributions such as Normal (or perhaps Laplace);

Tests for the error distribution

We wish to test for

- the specific parametric form of the error distribution F, i.e. whether or not it corresponds to a specific family of distributions such as Normal (or perhaps Laplace);

Why?

- justify the use of inferential tools (confidence intervals, $p-$values, etc.)
- check whether the model "fits well". If rejected, either of
  - the model
  - the method of estimation

  may be inadequate.

Characteristic functions

- Recall that, for $\varepsilon \sim F$, the characteristic function of ε is given by

$$\phi_F(t) = \mathbb{E}\left(e^{it\varepsilon}\right)$$

- Closed expressions exist for a wide range of distributions. For instance, if $F = N(0,1)$, then

$$\phi_{N(0,1)}(t) = e^{-\frac{1}{2}t^2}$$

- For observed residuals $\{\hat{\varepsilon}_i\}_{i=1}^n$, $\phi_F(t)$ can be estimated through the empirical characteristic function of the residuals,

$$\varphi_n(t) = \int e^{it\hat{\varepsilon}} dF_n(\hat{\varepsilon}) = \frac{1}{n}\sum_{j=1}^{n} e^{it\hat{\varepsilon}_j}$$

Testing for a parametric distribution

- $H_0 : \varepsilon \sim F_0$.
- Construct test by comparing $\varphi_n(t)$ with $\phi_{F_0}(t)$.
- Omnibus test statistic:

$$T = n \int_{-\infty}^{\infty} |\varphi_n(t) - \phi_{F_0}(t)|^2 w(t) dt,$$

- $w(t)$ is some weight function that is chosen so that T can be expressed in closed form.
- Specifically, if $F_0 \sim N(0,1)$, and $w(t) = e^{-at^2}$, then
$T \equiv T_a = \frac{1}{n}\sqrt{\frac{\pi}{a}} \left(\sum_{j,k=1}^{n} e^{-(\hat{\varepsilon}_j - \hat{\varepsilon}_k)^2/4a} \right) + n\sqrt{\frac{\pi}{1+a}} -$
$$2\sqrt{\frac{2\pi}{1+2a}} \left(\sum_{j=1}^{n} e^{-\frac{\hat{\varepsilon}_j^2}{(2+4a)}} \right).$$

Testing for a parametric distribution

- $H_0 : \varepsilon \sim F_0$.
- Construct test by comparing $\varphi_n(t)$ with $\phi_{F_0}(t)$.
- Omnibus test statistic:

$$T = n \int_{-\infty}^{\infty} |\varphi_n(t) - \phi_{F_0}(t)|^2 w(t) dt,$$

- $w(t)$ is some weight function that is chosen so that T can be expressed in closed form.
- Specifically, if $F_0 \sim N(0,1)$, and $w(t) = e^{-at^2}$, then
$T \equiv T_a = \frac{1}{n}\sqrt{\frac{\pi}{a}} \left(\sum_{j,k=1}^{n} e^{-(\hat{\varepsilon}_j - \hat{\varepsilon}_k)^2/4a} \right) + n\sqrt{\frac{\pi}{1+a}} -$
$$2\sqrt{\frac{2\pi}{1+2a}} \left(\sum_{j=1}^{n} e^{-\frac{\hat{\varepsilon}_j^2}{(2+4a)}} \right).$$
- The limiting distribution of T_a is unknown, and hard to derive.

Bootstrapped p-values

Reproduce the sampling distribution of T_a via the Bootstrap.

(i) On the basis of data $\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i\}$, compute estimators $(\hat{\boldsymbol{\beta}}, \hat{g}(\cdot), \hat{\sigma})$ and the corresponding residuals $\hat{\varepsilon}_i,\ \ i = 1, 2, ..., n$.

(ii) Compute the test statistic $T_a = T_a(\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n)$.

(iii) Repeat B times (typically, $B = 200$):

- Generate iid. replicates $\varepsilon_i^*, i = 1, 2, ..., n$, from F_0, and define bootstrap observations

$$y_i^* = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \hat{g}(\boldsymbol{z}_i) + \hat{\sigma}\varepsilon_i^*.$$

- Based on $\{y_i^*, \boldsymbol{x}_i, \boldsymbol{z}_i\}$, compute $(\hat{\boldsymbol{\beta}}^*, \hat{g}^*(\cdot), \hat{\sigma}^*)$ and corresponding residuals $\hat{\varepsilon}_i^*,\ \ i = 1, 2, ..., n$.
- Compute the test statistic $T_a^* := T_a(\hat{\varepsilon}_1^*, ..., \hat{\varepsilon}_n^*)$.

If k of the T_a^* exceed T_a, then $p = k/(B+1)$ is the $p-$value of the test.

Simulation study

Data sets of size $n = 100$ are generated from the model

$$y = x + \sin(2\pi z) + \sigma\varepsilon$$

where both x and z are uniformly distributed in the interval $[0, 1]$, and $\sigma = 0.5$. The simulated error distributions are:

(N) Gaussian distribution with mean 0 and standard deviation 1;

(L) Laplace distribution with mean 0 and scale parameter 1.

(SN) Skew-Normal distribution centered at 0, with scale parameter 1 and skew parameter 10;

(SL) Skew-Laplace distribution centered at 0, with scale parameter 1 and skew parameter 3.

Simulation study (cont.)

- We estimate the model, $y = \beta x + g(z) + \sigma\varepsilon$, using "backfitting" with a cubic spline smoother for g, which is calibrated to produce a nonparametric term corresponding to approximately 4 degrees of freedom.
- For each of the error distributions (N), (L), (SN), and (SL), we consider the null hypotheses $H_0^{(N)}$ and $H_0^{(L)}$, i.e. Normal and Laplace–distributed error, respectively.
- We generate 2000 Monte replications for each test problem and count the number of rejections of the corresponding null hypothesis.
- The test is compared with the (bootstrapped versions of the) classical Cramér–von Mises (CM) and Anderson–Darling (AD) tests, which employ empirical distribution functions (rather than empirical characteristic functions).

Simulation study (cont.)

- Percentage of rejection of the null hypothesis $H_0^{(N)}$ (Normality) for four different true error distributions.

		$a = 1/2$	$a = 1$	$a = 2$	AD	CM
(N)	$\alpha = 0.05$	5.0	5.1	5.0	4.3	4.2
	$\alpha = 0.10$	9.2	10.0	10.1	9.0	8.8
(L)	$\alpha = 0.05$	77.9	75.7	70.5	71.1	69.4
	$\alpha = 0.10$	86.5	84.7	81.5	80.1	79.5
(SN)	$\alpha = 0.05$	84.5	85.6	85.3	82.6	76.7
	$\alpha = 0.10$	90.8	91.6	91.0	89.1	85.3
(SL)	$\alpha = 0.05$	100.0	100.0	100.0	100.0	99.9
	$\alpha = 0.10$	100.0	100.0	100.0	100.0	100.0

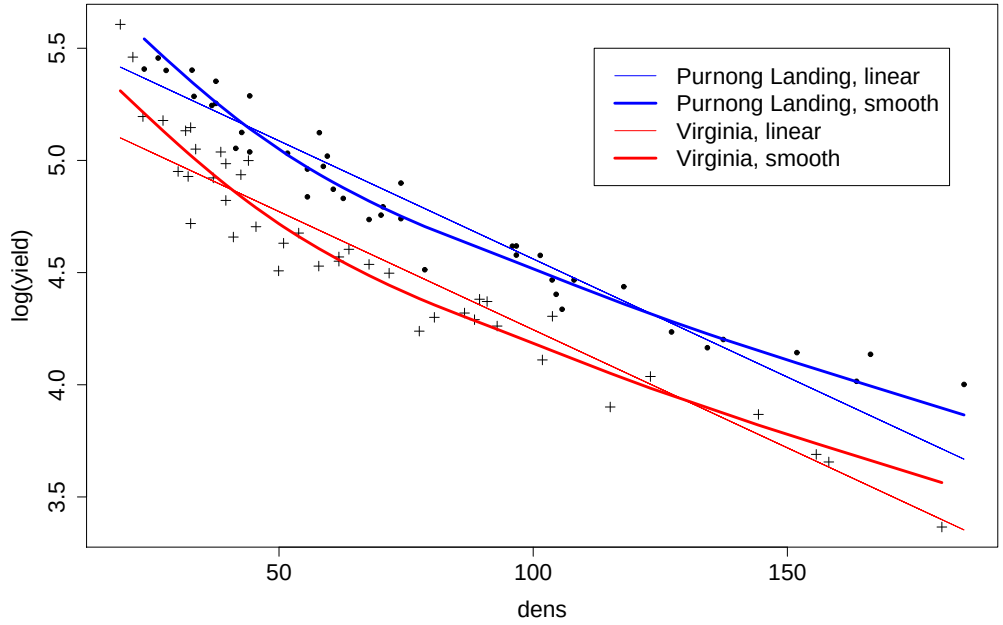
Simulation study (cont.)

- Percentage of rejection of the null hypothesis $H_0^{(L)}$ (Laplace) for four different true error distributions.

		$a = 1/2$	$a = 1$	$a = 2$
(N)	$\alpha = 0.05$	57.9	62.0	44.3
	$\alpha = 0.10$	73.1	75.5	65.5
(L)	$\alpha = 0.05$	4.8	4.9	4.6
	$\alpha = 0.10$	11.2	10.4	9.7
(SN)	$\alpha = 0.05$	91.0	95.6	93.9
	$\alpha = 0.10$	95.6	97.9	98.0
(SL)	$\alpha = 0.05$	100.0	99.9	99.9
	$\alpha = 0.10$	100.0	100.0	100.0

Back to onions data

- We fit both the linear and the semi-linear model, and compute the $p-$value for testing H_0: Normality of errors.



- One obtains
  - for the linear model, $p = 0.00$ (Normality is rejected).
  - for the semi–linear model, $p = 0.42$ (Normality is *not* rejected).
- We conclude that the semi-linear model fits significantly better than the linear model.

Related test problems

- Symmetry test: H_0: F is symmetric.
  - Key idea: Decompose characteristic function of ε,

$$\phi_F(t) = \mathbb{E}(\cos(t\varepsilon)) + i\mathbb{E}(\sin(t\varepsilon)) \equiv C(t) + iS(t)$$

  - $C(t)$ captures the full information on the symmetric component of the error distribution.
  - Hence, the Fourier formulation of H_0 is

$$H_0 : S(t) = 0, \quad t \in \mathbb{R},$$

  and we use the test statistic

$$S = n \int S_n^2(t) w(t)\, dt$$

  with $S_n(t) = \frac{1}{n} \sum_{i=1}^{n} \sin(t\hat{\varepsilon}_i)$

Related test problems (cont.)

- Model specification test: Given covariates $\boldsymbol{v} = (\boldsymbol{x}', \boldsymbol{z}')'$, test

$$H_0 : y = \boldsymbol{x}'\boldsymbol{\beta} + g(\boldsymbol{z}) + \varepsilon$$

for some $\boldsymbol{\beta} \in \mathbb{R}^p$, $g : \mathbb{R}^p \longrightarrow \mathbb{R}$.

  - Key idea (adapted from Bierens, 1982): H_0 is true iff

$$\mathbb{E}[\{y - \boldsymbol{x}'\boldsymbol{\beta} - g(\boldsymbol{z})\}e^{i\boldsymbol{t}'\boldsymbol{v}}] = 0, \ \forall \ \boldsymbol{t} \in \mathbb{R}^{p+q}, \tag{1}$$

    which we estimate via $E_n(\boldsymbol{t}) = \frac{1}{n}\sum_{i=1}^n \widehat{\varepsilon}_i e^{i\boldsymbol{t}'\boldsymbol{v}_i}$.
  - Omnibus procedure for specification testing is to reject the null hypothesis H_0 for large values of the test statistic

$$R = \int_{\mathbb{R}^{p+q}} |E_n(\boldsymbol{t})|^2 \, w(\boldsymbol{t}) d\boldsymbol{t}, \tag{2}$$

Related test problems (cont.)

- For these tests, the bootstrap routines are not identical to that one given earlier, but similar in spirit, see Meintanis & Einbeck (2010, 2011).
- Results for onions data:
- Test of H_0: F is symmetric
  - LM: $p = 0.12$ (close to rejection)
  - Semi-LM: $p = 0.69$ (clearly not rejected).
- Specification test for

$$H_0 : \mathbb{E}(\log(\texttt{yield}|\texttt{location, dens}) = \beta_0 + \beta_1 \texttt{location} + \ldots$$

  - $\ldots \beta_2 \texttt{dens}$: $p = 0.07$ (LM rejected at the 10% level).
  - $\ldots g(\texttt{dens})$: $p = 0.31$ (Semi-LM not rejected).

Conclusion

- Empirical characteristic functions are a versatile tool for tackling a wide range of test problems in the context of semi- and non-parametric regression.
- The limit distribution of the resulting test statistics is difficult to derive, except for special cases as the linear model, or the univariate Nadaraya-Watson estimator (Hušková & Meintanis, 2007, 2010).
- However, suitably adapted bootstrap routines can be conveniently employed instead.
- The methods achieve generally higher test powers than tests based on the empirical distribution function (whether these are bootstrapped or not), and good compliance with the target significance level.

References

Bierens, H.J. (1982). Consistent model specification tests. *Econometrica* **20**, 105–134.

Hušková, S., and & Meintanis, S. (2007). Omnibus tests for the error distribution linear regression models. *Statistics* **41**, 363–376.

Hušková, S., and & Meintanis, S. (2010). Tests for the error distribution in nonparametric possibly heteroscedastic regression models. *Test* **19**, 92–112.

Meintanis, S., and Einbeck, J. (2010). Goodness-of-fit tests in semi-linear models. Preprint, submitted, under review.

Meintanis, S., and Einbeck, J. (2011). Validation tests for semiparametric models. Working paper.