Chapter 4

Forecast Model Building and Use

It has been said that "It's difficult to make predictions, especially about the future." This quip is funny insofar as *all* predictions are about the future. But actually they're not. Prediction is a major topic even in cross-sections, in which there is no temporal aspect. In this chapter we consider cross-section prediction.

4.1 Cross-Section Prediction

The environment is:

$$y_i = x_i'\beta + \varepsilon_i,\ i = 1, ..., N$$

$$\varepsilon_i \sim\ iid\ D(0, \sigma^2).$$

In cross sections, everything is easy. That is, cross-section prediction simply requires evaluating the conditional expectation (regression relationship) at a *chosen* value of x, $x = x^*$. Suppose, for example, that we know a regression relationship between expenditure on restaurant meals (y) to income (x). If we get new income data for a new household, we can use it to predict its restaurant expenditures.

4.1.1 Point Prediction

Continue to assume for the moment that we know the model parameters. That is, assume that we know β and all parameters governing D.[1]

We immediately obtain point forecasts as:

$$E(y_i|x_i = x^*) = x^{*\prime}\beta.$$

4.1.2 Density Prediction for D Gaussian

Density forecasts, and hence interval forecasts, are a bit more involved, depending on what we're willing to assume. In any event the key is somehow to account for **disturbance uncertainty**, the part of forecast uncertainty arising from the fact that our forecasting models involve stochastic disturbances.

If D is Gaussian, then the density prediction is immediately

$$y_i|x_i = x^* \ \sim \ N(x^{*\prime}\beta, \sigma^2). \tag{4.1}$$

We can calculate any desired interval forecast from the density forecast. For example, a 95% interval would be $x^{*\prime}\beta \pm 1.96\sigma$.

Now let's calculate the density and interval forecasts by a more roundabout simulation algorithm that will be very useful in more complicated (and realistic) cases.

1. Take R draws from the disturbance density $N(0, \sigma^2)$.
2. Add $x^{*\prime}\beta$ to each disturbance draw.
3. Form a density forecast by fitting a density to the output from step 2.
4. Form an interval forecast (95%, say) by sorting the output from step 2 to get the empirical cdf, and taking the left and right interval endpoints

[1] Note that the mean and variance are in general insufficient to characterize a non-Gaussian D.

as the the .025% and .975% values, respectively.

As $R \to \infty$, the algorithmic results coincide with those of 4.1.

4.1.3 Density Prediction for D Parametric Non-Gaussian

Our simulation algorithm still works for non-Gaussian D, so long as we can simulate from D.

1. Take R draws from the disturbance density D.
2. Add $x^{*\prime}\beta$ to each disturbance draw.
3. Form a density forecast by fitting a density to the output from step 2.
4. Form a 95% interval forecast by sorting the output from step 2, and taking the left and right interval endpoints as the the .025% and .975% values, respectively.[2]

Again as $R \to \infty$, the algorithmic results become arbitrarily accurate.

4.1.4 Making the Forecasts Feasible

The approaches above are infeasible in that they assume known parameters. They can be made feasible by replacing unknown parameters with estimates. For example, the feasible version of the point prediction is $x^{*\prime}\hat{\beta}$. Similarly, to construct a feasible 95% interval forecast in the Gaussian case we can take $x^{*\prime}\hat{\beta} \pm 1.96\hat{\sigma}$, where $\hat{\sigma}$ is the standard error of the regression (also earlier denoted s).

[2]Note that, now that we have in general abandoned symmetry, the prescribed method no longer necessarily generates the shortest interval.

4.1.5 Density Prediction for D Non-Parametric

Now assume that we know nothing about distribution D, except that it has mean 0. In addition, now that we have introduced "feasible" forecasts, we will stay in that world.

1. Take R draws from the regression residual density (which is an approximation to the disturbance density) by assigning probability $1/N$ to each regression residual and sampling with replacement.
2. Add $x^{*\prime}\hat{\beta}$ to each draw.
3. Form a density forecast by fitting a density to the output from step 2.
4. Form a 95% interval forecast by sorting the output from step 2, and taking the left and right interval endpoints as the the .025% and .975% values, respectively.

As $R \to \infty$ and $N \to \infty$, the algorithmic results become arbitrarily accurate.

4.1.6 Density Forecasts for D Nonparametric and Acknowledging Parameter Estimation Uncertainty

Thus far we have accounted only for disturbance uncertainty in our feasible density forecasts. Disturbance uncertainty reflects the fact that disturbance realizations are inherently unpredictable. There is simply nothing that we can do about disturbance uncertainty; it is present always and everywhere, even if we were somehow to know the DGP and its parameters.

We now consider **parameter estimation uncertainty**. The coefficients that we use to produce predictions are of course just *estimates*. That is, even if we somehow know the form of the DGP, we still have to estimate its parameters. Those estimates are subject to sampling variability, which makes

an additional contribution to prediction errors. The "feasible" approach to density forecasting sketched above still fails to acknowledge parameter estimation uncertainty, because it treats "plugged-in" parameter estimates as true values, ignoring the fact that they are only estimates and hence subject to sampling variability. Parameter estimation uncertainty is often ignored, as its contribution to overall forecast MSE can be shown to vanish unusually quickly as sample size grows (See EPC 1). But it impacts forecast uncertainty in small samples and hence should not be ignored in general.

1. Take R approximate disturbance draws by assigning probability $1/N$ to each regression residual and sampling with replacement.

2. Take R draws from the large-N sampling density of $\hat{\beta}$, namely

$$\hat{\beta}_{OLS} \sim N(\beta, \sigma^2(X'X)^{-1}),$$

   as approximated by $N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$.

3. To each disturbance draw from 1 add the corresponding $x^{*\prime}\hat{\beta}$ draw from 2.

4. Form a density forecast by fitting a density to the output from step 3.

5. Form a 95% interval forecast by sorting the output from step 3, and taking the left and right interval endpoints as the the .025% and .975% values, respectively.

As $R \to \infty$ and $N \to \infty$, we get precisely correct results.

4.1.7 Incorporating Heteroskedasticity

We will illustrate for the Gaussian case without parameter estimation uncertainty, using an approach that closely parallel's White's test for heteroskedas-

ticity. Recall the feasible density forecast,

$$y_i|x_i = x^* \sim N(x^{*\prime}\hat{\beta}, \hat{\sigma}^2).$$

Now we want to allow for the possibility that $\hat{\sigma}^2$ varies with x_i.

1. Regress by OLS: $y_i \to x_i$ and save the residuals e_i.
2. Regress $e_i^2 \to x_i$. Call the estimated coefficient vector $\hat{\gamma}$.
3. Form the density forecast as

$$y_i|x_i = x^* \sim N(x^{*\prime}\hat{\beta}, \hat{\sigma}^2(x^*)),$$

where $\hat{\sigma}^2(x^*) = x^{*\prime}\hat{\gamma}$ is the fitted value from the regression in step 2 evaluated at x^*.

One could of course run regression 1 by weighted least squares (WLS) rather than OLS using the $\hat{\sigma}^2(x^*)$ as weights, but the efficiency gains in estimating β are not likely to produce large additional improvements in calibration of density and interval forecasts. The key is to allow the disturbance variance to adapt to x^* when forming forecasts, quite apart from whether they are centered at $x^{*\prime}\hat{\beta}_{OLS}$ or $x^{*\prime}\hat{\beta}_{WLS}$.

4.2 Wage Prediction Conditional on Education and Experience

4.2.1 The CPS Dataset

We will examine the CPS wage dataset, containing data on a large cross section of individuals on wages, education, experience, sex, race and union status. For a detailed description see Appendix E. For now we will use only wage, education and experience.

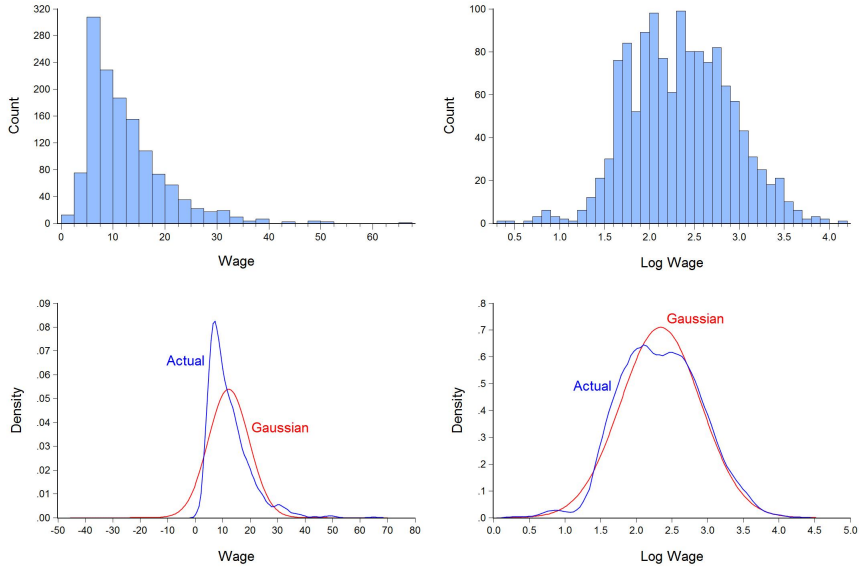


Figure 4.1: Distributions of Wages and Log Wages

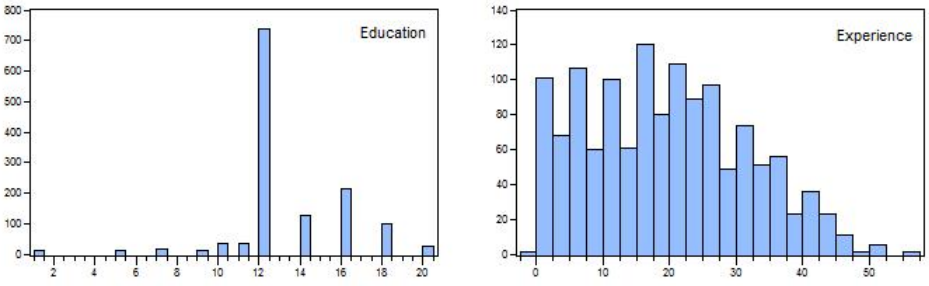


Figure 4.2: Histograms for Wage Covariates

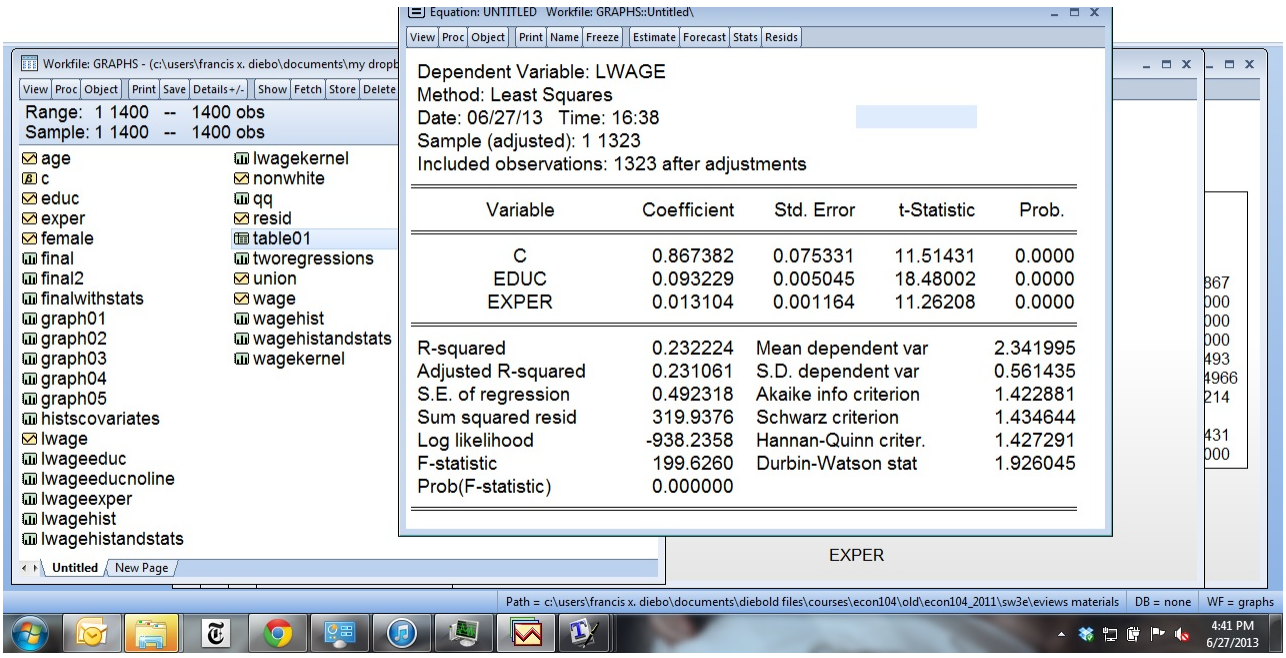


Figure 4.3: Wage Regression on Education and Experience

– Basic features of wage, education and experience data.

In Figures 4.1 and 4.2 we show histograms and statistics for potential determinants of wages. Education (EDUC) and experience (EXPER) are standard continuous variables, although we measure them only discretely (in years).

4.2.2 Regression

– Linear regression of log wage on predictors (education and experience).

Recall our basic wage regression,

$$LWAGE \rightarrow c, EDUC, EXPER,$$

shown in Figure 4.3. Both explanatory variables are highly significant, with expected signs. In table **??** consider a linear versus a quadratic model.

Even though the quadratic regression coefficients are statistically significant, we see only an extremely small improvement in adj. R^2 and RMSE. We

also consider the histograms for the two models, in Figure **??**.

We can see that the densities of residuals are almost identical, perhaps that those from the quadratic model are ever so slightly less skewed. Since we believe in the parsimony principle however, we will restrict ourselves to a linear model in the absence of overwhelming evidence in favor of a non-linear model. NOTE: There are many more nonlinear models to try besides quadratic! See section 4.3 for possible further extensions.

Throughout we will use the "best" estimated log wage model for feasible prediction of wage, for (EDUC, EXPER) = (10, 1) and (EDUC, EXPER) = (14, 20). (NOTE: The model is for log wage, but the forecasts are for wage.)

4.2.3 Point Prediction by Exponentiating vs. Simulation

An obvious point forecast of $WAGE$ can be obtained by exponentiating a forecast of $LWAGE$. But there are issues. In particular, if $lny_{t+h,t}$ is an unbiased forecast of lny_{t+h}, then $\exp(lny_{t+h,t})$ is a *biased* forecast of y_{t+h}.[3] More generally, if $(f(y))_{t+h,t}$ is an unbiased forecast of $(f(y))_{t+h}$, then $f^{-1}((f(y))_{t+h,t})$ is a biased forecast of y_{t+h}, for arbitrary nonlinear function f, because the expected value of a nonlinear function of a random variable does not equal the nonlinear function of the expected value, a result known as Jensen's inequality.[4]

Various analytic "bias corrections" have been proposed, but they rely on strong and unnecessary assumptions. The modern approach is simulation-based. Using simulation, simply build up the density forecast of the object of interest (e.g., $WAGE$ rather than $lnWAGE$), the sample mean of which across simulations is consistent for the population conditional mean. The bias correction is done automatically!

[3] A forecast is unbiased if its mean error is zero. Other things equal, unbiasedness is desirable.

[4] As the predictive regression $R^2 \to 1$, however, the bias vanishes. Why?

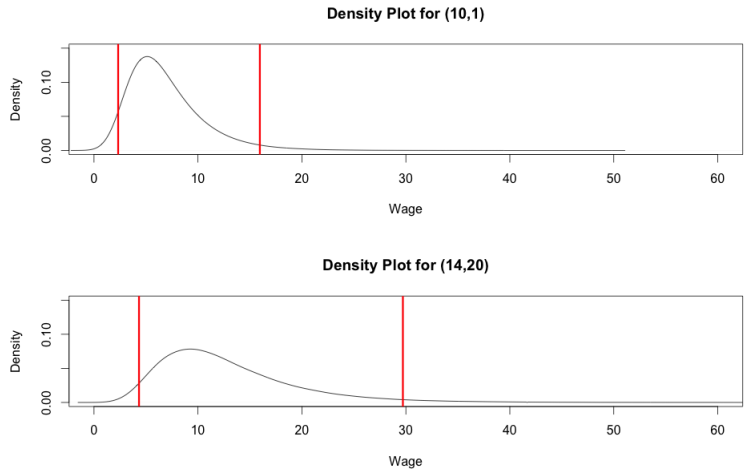


Figure 4.4: Predicted densities for wage under the assumption that residuals are homoskedastic and Gaussian, abstracting from parameter uncertainty. The model is in logs and then exponentiated The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

4.2.4 Density Prediction for D Gaussian

We now apply the methods from section 4.1.2 to the linear model. To operationalize that algorithm, we must first make an estimator of σ^2 and β. $\hat{\beta}$ is taken directly as the OLS regression coefficients, and $\hat{\sigma}$ can be taken as the residual standard error. With those plug-in estimators found, we can follow the algorithm directly. Since we are in a Gaussian environment, recall we could find a 95% CI by taking $x^{*\prime}\beta \pm 1.96\hat{\sigma}$. However, in more complex environments, we will have to take the CI directly from the simulated data, so we will do that here by sorting the sample draws and taking the left and right endpoints to be the .025% and .975% values, respectively. This yields the output from figure 4.4.

Two things are of particular note here. First is that, as expected, the density prediction for the individual with more education and experience has a much higher mean. Second is that the CI for individual 2 is much wider than that of individual 1, or similarly that the density prediction has much higher variance. This is perhaps surprising, since we were in a case with

homoskedasticity. In fact this is one of the costs of working with a log-linear model for wages:

$$\begin{aligned} \log(y) &= x'\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \\ y &= \exp(x'\beta)\exp(\varepsilon) \\ \Rightarrow \mathbb{V}[y] &= [\exp(x'\beta)]^2 \mathbb{V}[\exp(\varepsilon)] \end{aligned}$$

Thus even with homoskedasticity in logs, the variance of the level y_t will depend on x.

4.2.5 Density Forecasts for D Gaussian and Acknowledging Parameter Estimation Uncertainty

We are still in a sufficiently simple world that we may follow directly the algorithm above. A quick way to think about the algorithm of the previous section is the following: Since residuals are Gaussian, y is Gaussian. So to compute a density prediction of y, all we really need is to estimate its mean and covariance. The mean is given directly as the conditional mean from the model. For the covariance:

$$\begin{aligned} \mathbb{V}[y] &= \mathbb{V}[x'\beta + \varepsilon] \\ &= \mathbb{V}[x'\beta] + \mathbb{V}[\varepsilon] \end{aligned}$$

Since the previous section did not allow for parameter estimation uncertainty, the first term in that sum was zero. We will now accurately estimate that first term and include it in our density prediction. This idea is explored more in the EPC's.

Having Gaussian disturbances means that the distribution of $\hat{\beta}$ is precisely

Gaussian, and as above we knows its mean and covariance: β and $\sigma^2(X'X)^{-1}$. To operationalize this, we will make draws from $N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$. Concretely, our algorithm is the following:

1. Take R draws from the estimated disturbance density $N(0, \hat{\sigma}^2)$.
2. Take R draws of β from the estimated parameter sampling distribution $N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$.
3. Add the disturbance draw from step 1 to the draw of $x^{*'}\beta$, where β is drawn as in step 2.
4. Exponentiate each draw to turn the draw of log wage into a draw for wage.
5. Form the density forecast by fitting a density to the output.
6. Form a 95% interval forecast by sorting the output, and taking the left and right interval endpoints as the .025% and .975% values, respectively.

Following this algorithm yields the output of figure 4.5. We see that these density forecasts are nearly identical to those without parameter uncertainty. This is to be expected once we consider the estimated covariance matrix of β, which we find has very small variance:

$$\mathbb{V}[\hat{\beta}] = \begin{bmatrix} 0.00567 & -0.000357 & -4.19e\text{-}05 \\ -0.000357 & 2.55e\text{-}05 & 1.22e\text{-}06 \\ -4.19e\text{-}05 & 1.22e\text{-}06 & 1.35e\text{-}06 \end{bmatrix}$$

4.2.6 Density Forecasts for D Gaussian, Acknowledging Parameter Estimation Uncertainty, and Allowing for Heteroskedasticity

For this section we find that we must work a little bit harder. There are two separate difficulties that are important to get correct: The first is an

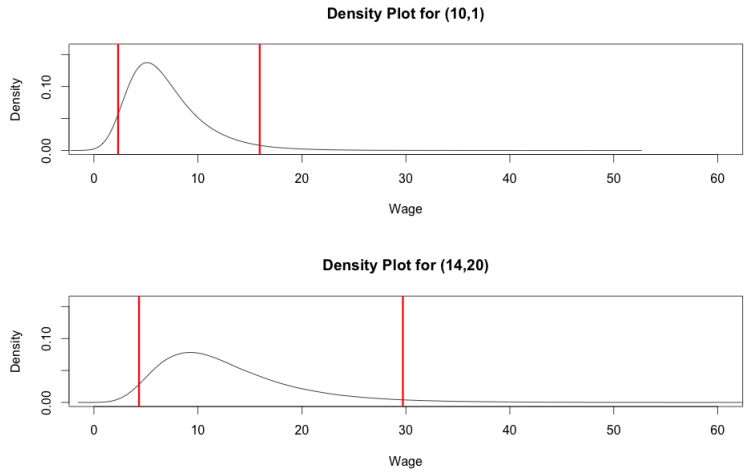


Figure 4.5: Predicted densities for wage under the assumption that residuals are homoskedastic and Gaussian. Here parameter uncertainty is accounted for in the density of wage. The model is in log wage and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

appropriate drawing of the sampled residuals, the second is an appropriate drawing of the parameters.

First, the parameter estimation uncertainty. The covariance matrix we estimated above, $\sigma^2(X'X)^{-1}$, is no longer valid in the presence of heteroskedasticity. Rather:

$$\begin{aligned} \hat{\beta}_{OLS} &= (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon \\ &\Rightarrow \hat{\beta} \sim N(\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1}) \end{aligned} \tag{4.2}$$

Under homoskedasticity, $\Omega = \sigma^2 I$ and so this covariance matrix dramatically simplifies. This is no longer the case under heteroskedasticity. In this environment we will find the distinction to be of little numerical importance, but for other datasets it will be of dramatic importance.

Of course, in the presence of heteroskedasticity, we may prefer to instead conduct weighted least squares instead of OLS. Recall the WLS estimator is

$$\hat{\beta}_{WLS} = (X'\Sigma X)^{-1}X'\Sigma Y$$

Here Σ is any diagonal weighting matrix. A popular choice is of course Ω^{-1}, as this choice is efficient, where this matrix can be estimated by a number of two-stage processes. The asymptotic covariance matrix of the WLS estimator is then

$$\begin{aligned} & \hat{\beta}_{WLS} = \beta + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\varepsilon \\ \Rightarrow & \hat{\beta}_{WLS} \sim N(\beta, (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1})^{-1}) \\ \Rightarrow & \hat{\beta}_{WLS} \sim N(\beta, (X'\Omega^{-1}X)^{-1}) \end{aligned} \tag{4.3}$$

Thus β_{WLS} is simultaneously a better estimator than $\hat{\beta}_{OLS}$ and with an easier covariance matrix to estimate. For this reason we will proceed using $\hat{\beta}_{WLS}$, and make draws from the above. To do this, we must select a specific two-stage process, as discussed above. We will discuss this in the course of the estimation of the residual density. This is done via the following algorithm:

1. Regress by OLS: $y_i \rightarrow x_i$ and save the residuals.
2. Regress $e_i^2 \rightarrow x_i$. Call the estimated coefficient vector $\hat{\gamma}$.
3. Construct the vector of heteroskedasticities $\hat{\sigma}^2(x_i) = x_i'\hat{\gamma}$, and set $\hat{\Omega} = diag(\hat{\sigma}^2(x_i))$.
   - Use $\hat{\Omega}$ to conduct WLS regression.
4. Take R draws of the residuals from $N(0, \hat{\sigma}^2(x^*))$
5. Take R draws of β from $N(\hat{\beta}, (X'\hat{\Omega}^{-1}X)^{-1})$
6. Add the disturbance draw from step 4 to the draw of $x^{*'}\beta$, where β is drawn as in step 5.
7. Exponentiate each draw to turn the draw of log wage into a draw for wage.
8. Form the density forecast by fitting a density to the output.

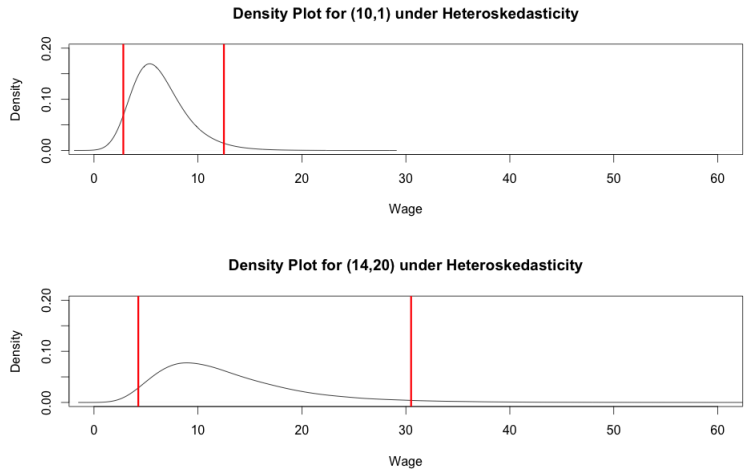


Figure 4.6: Predicted densities for wage under the assumption that residuals are heteroskedastic and Gaussian. Here parameter uncertainty is accounted for in the density of wage. The model is in log wage and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

9. Form a 95% interval forecast by sorting the output, and taking the left and right interval endpoints as the .025% and .975% values, respectively.

Notice one could argue with our above procedure: We took residuals from the OLS regression to make the density prediction. There is certainly an argument to be made from re-taking residuals from the WLS regression and re-estimating the heteroskedasticity covariance matrix from there. However, the above will still be a consistent procedure (since the covariance matrix estimated is HAC-consistent), and since we have a surplus of observations we are unlikely to see a numerical difference between the two. The outputs from this procedure can be found in figure 4.6.

The CI interval for the lower wage, lower educated individual shrunk dramatically, while the CI for the $(14, 20)$ individual did not change noticeably. This is explainable by the fact that the average number of years of education for our dataset is 13.1, and the average number of years of experience is 19.2. Thus the $(14, 20)$ individual is close to the average. Since the form of heteroskedasticity is measured to be linear in this algorithm, and homoskedas-

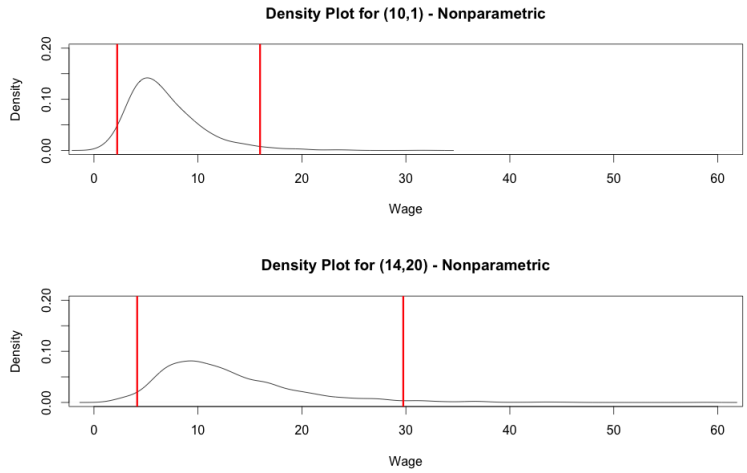


Figure 4.7: Predicted densities for wage under the assumption that residuals are homoskedastic, abstracting from parameter uncertainty. The density of residuals is now estimated nonparametrically. The model is in log wages and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

ticity will set the variance of each individual to be approximately the mean variance of the dataset, it is expected that the mean individual will have approximately the same variance under homoskedasticity and heteroskedasticity.

4.2.7 Density Prediction for D Nonparametric

In this section we will make density predictions for our dataset dropping the assumption that disturbances are Gaussian. For now we will assume that we can estimate parameters with no uncertainty and that disturbances are homoskedastic. Here we may follow the exact algorithm of 4.1.5, with the added step that we exponentiate each draw to make a draw for wage from a draw for log wage. The yielded output can be found in 4.7.

Here we find that the nonparametric density estimates are quite similar to those found assuming D Gaussian. This suggests that our assumption of Gaussian disturbances was well-grounded. We can examine this further by

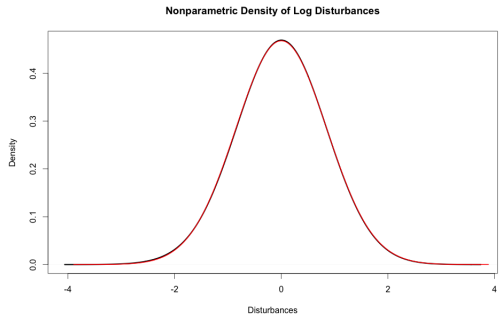


Figure 4.8: Nonparametric Density of Disturbances. Red overlaid line is Gaussian.

directly observing the nonparametric density of the disturbances to the log wages, seen in figure 4.8. These are some very Gaussian disturbances.

4.2.8 Density Forecasts for D Nonparametric and Acknowledging Parameter Estimation Uncertainty

Here we will blend the algorithms of the previous sections. Even though having non-Gaussian disturbances no longer assures that $\hat{\beta}$ is precisely Gaussian, by CLT the normal distribution remains the large-N approximation. Thus, we may construct the algorithm exactly as in section 4.1.6, as before with the added step of exponentiating each log wage draw to get a draw for wage. This yields the output in figure 4.9.

Comparing the results from nonparametric estimation, and the results from just incorporating parameter estimation uncertainty, we should be unsurprised by this: nonparametric estimation told us that our Gaussian disturbances assumption was well-grounded, and our initial parameter uncertainty estimation results told us our parameter estimates were being measured quite accurately. Thus the output from this section very closely resembles that of our very first density predictions, with D Gaussian and no parameter uncertainty.

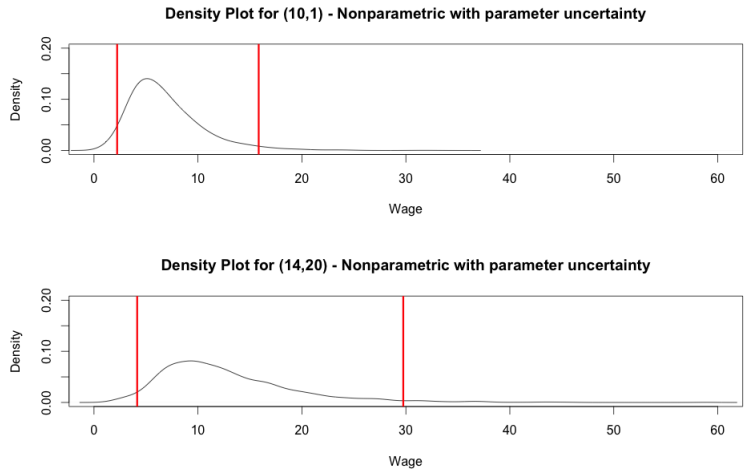


Figure 4.9: Predicted densities for wage under the assumption that residuals are homoskedastic. Here parameter uncertainty is accounted for in the density of wage, and the density of residuals is now estimated nonparametrically. The model is in log wage and then exponentiated. The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

4.2.9 Modeling Directly in Levels

Up until now our model has been

$$\log(y) = x'\beta + \varepsilon$$

From this model we construct density predictions for y by making draws of $\log(y)$ and exponentiating. We now switch to the following model:

$$y = x'\beta + \varepsilon$$

We will now re-explore the above analysis in this context. The first thing we will notice is that density predictions under the assumption of Gaussian disturbances will generally perform quite poorly, because the disturbances to the level model are quite clearly non-Gaussian. See figure 4.10.

We therefore skip to a nonparametric density prediction, incorporating parameter uncertainty - although as before we will find that parameter un-

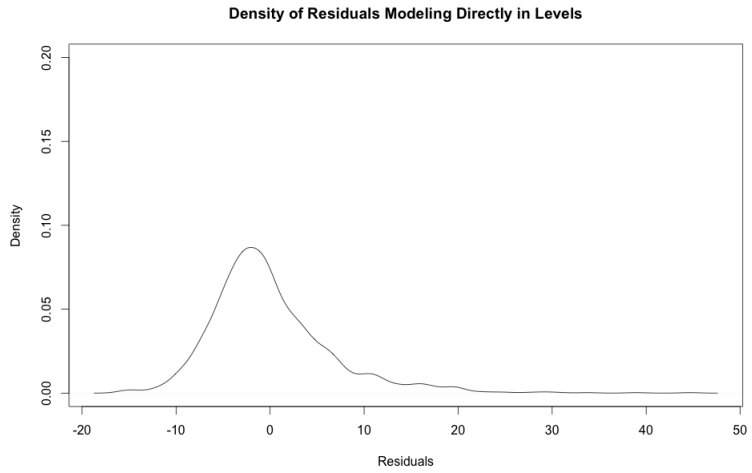


Figure 4.10: Non-Gaussian Residuals

certainty is quite small. The resulting output is in figure 4.11

We immediately notice a problem for individual 1: The 95% CI includes negative values for wage! This is inherently a problem of working directly in levels when modeling a variable for which only positive values make sense.

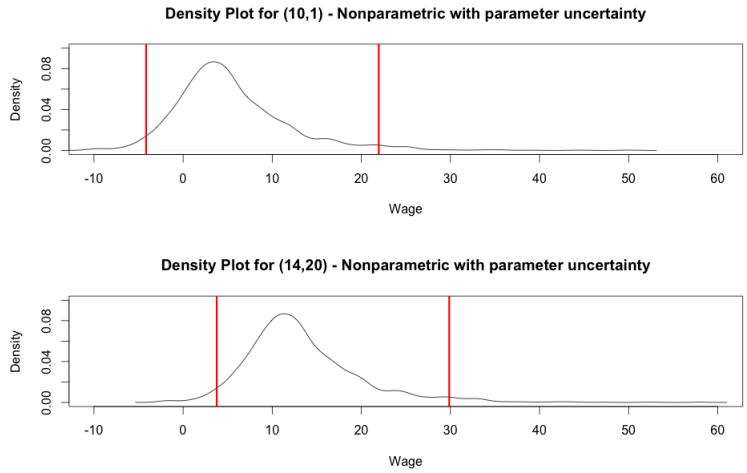


Figure 4.11: Predicted densities for wage under the assumption that residuals are homoskedastic. Here parameter uncertainty is accounted for in the density of wage, and the density of residuals is now estimated nonparametrically. The model is directly in wages The top graph is estimated density of wage with 10 years of education, 1 year of experience. The bottom graph is the same for 14 years of education, 20 years of experience. The red vertical lines indicate 95% CI.

4.3 Non-Parametric Estimation of Conditional Mean Functions

4.3.1 Global Nonparametric Regression: Series

In the bivariate case we can think of the relationship as

$$y_t = g(x_t, \varepsilon_t),$$

or slightly less generally as

$$y_t = f(x_t) + \varepsilon_t.$$

First consider **Taylor series expansions** of $f(x_t)$. The linear (first-order) approximation is

$$f(x_t) \approx \beta_1 + \beta_2 x,$$

and the quadratic (second-order) approximation is

$$f(x_t) \approx \beta_1 + \beta_2 x_t + \beta_3 x_t^2.$$

In the multiple regression case, the Taylor approximations also involves interaction terms. Consider, for example, $f(x_t, z_t)$:

$$f(x_t, z_t) \approx \beta_1 + \beta_2 x_t + \beta_3 z_t + \beta_4 x_t^2 + \beta_5 z_t^2 + \beta_6 x_t z_t + ...$$

Such **interaction effects** are also relevant in situations involving dummy variables. There we capture interactions by including products of dummies.[5]

Now consider **Fourier series expansions**. We have

$$f(x_t) \approx \beta_1 + \beta_2 sin(x_t) + \beta_3 cos(x_t) + \beta_4 sin(2x_t) + \beta_5 cos(2x_t) + ...$$

One can also mix Taylor and Fourier approximations by regressing not only on powers and cross products ("Taylor terms"), but also on various sines and cosines ("Fourier terms"). Mixing may facilitate parsimony.

The ultimate point is that so-called "intrinsically non-linear" models are themselves linear when viewed from the series-expansion perspective. In principle, of course, an infinite number of series terms are required, but in practice nonlinearity is often quite gentle so that only a few series terms are required (e.g., quadratic).

The Curse of Dimensionality

Let p be the adopted expansion order. Things quickly get out of hand as p grows, for fixed N.

[5]Notice that a product of dummies is one if and only if both individual dummies are one.

Bandwidth Selection and the Bias-Variance Tradeoff

For fixed N, smaller p reduces variance but increases bias, larger p reduces bias but inflates variance.

Good things happen as $p \to \infty$ while $p/N \to 0$.

p can be chosen by any of the criteria introduced earlier.

4.3.2 Local Nonparametric Regression: Nearest-Neighbor

Here we introduce the idea of local regression based on "nearest neighbors". It is a leading example of a local smoother. The basic model is

$$y_t = g(x_t) + \varepsilon_t.$$

Unweighted Locally-Constant Regression

We want to fit (predict) y for an arbitrary x^*. We use the x variables in a neighborhood of x^*, $n(x^*)$. In particular we use the P_T nearest neighbors. P_T can be chosen by CV. We find the P_T nearest neighbors using the Euclidean norm:

$$\lambda(x^*, x^*_{P_N}) = [\Sigma_{k=1}^{K}(x^*_{P_N k} - x^*_k)^2]^{\frac{1}{2}}.$$

The fitted value is then

$$\widehat{y}(x^*) = \frac{1}{P_N} \sum_{j \in n(x^*)} y_j$$

This "nearest-neighbor" idea is not only simple, but tremendously important for prediction. If we want to predict y for an arbitrary x^*, it is natural to examine and average the y's that happened for close x's.

Weighted Locally-Linear Regression

We will use the "tri-cube" neighborhood weight function:

$$v_i(x_i, x^*, x^*_{P_N}) = C\left(\frac{\lambda(x_i, x^*)}{\lambda(x^*, x^*_{P_N})}\right),$$

where

$$C(u) = \begin{cases} (1-u^3)^3 & for\ u < 1 \\ 0 & otherwise \end{cases}$$

We then obtain the fitted value by weighted linear regression:

$$\hat{y}^* = \hat{g}(x^*) = x^{*\prime}\hat{\beta}$$

where

$$\hat{\beta} = argmin[\Sigma_{i=1}^N v_i(y_i - x_i'\beta)^2]$$

Good things happen as $P_N \to \infty$ while $P_N/N \to 0$.

Figure 4.12: Locally Weighted Regression

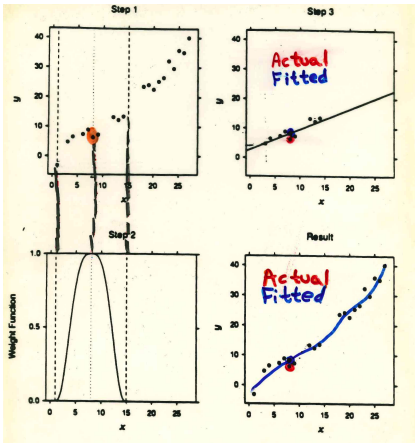
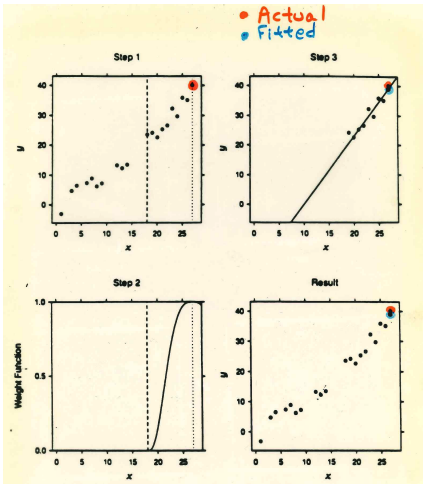


Figure 4.13: Locally Weighted Regression, Near the Edge



"Robustness Iterations"

Consider the initial fit to be "robusness iteration 0". The define the robustness weight at iteration 1:

$$\rho_i{}^{(1)} = S\left(\frac{e_i{}^{(0)}}{6h}\right)$$

where

$$e_i{}^{(0)} = y_i - \hat{y}_i^{(0)}$$

$$h = med\,|e_i{}^{(0)}|$$

$$S(u) = \begin{cases} (1-u^2)^2 & for |u| < 1 \\ 0 & otherwise \end{cases}$$

That is, we use bi-square robustness weighting, with bigger observations with bigger absolute residuals at iteration (0) downweighted progressively more, and observations with absolute residuals greater than six times the median absolute residual completely eliminated. We then obtain the fitted value by doubly-weighted linear regression:

$$\hat{y}_i^{*(1)} = \hat{g}^{(1)}(x^*) = x^{*\prime}\hat{\beta}^{(1)}$$

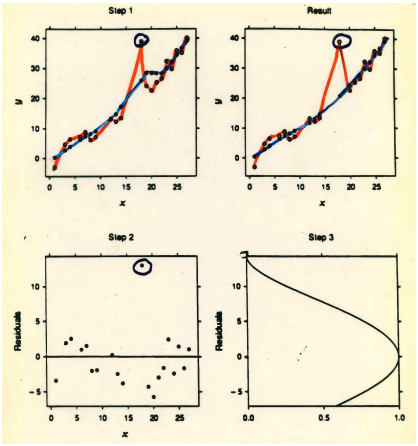
where

$$\hat{\beta}^{(1)} = argmin[\Sigma_{i=1}^{N} \rho_i^{(1)} v_i (y_i - x_i'\beta)^2].$$

Then we continue iterating if desired.

We speak of "robust weighted locally-linear regression". Extensions to locally-polynomial regression are immediate.

Figure 4.14: Locally Weighted Regression, Robustness Weighting for Outliers



4.3.3 Forecasting Perspectives

On Global vs. Local Smoothers for Forecasting

In cross-section environments, both global and local smoothers are useful for prediction. Local smoothers are perhaps more flexible and more popular in cross sections. x^* is usually interior to the observed x's, so nearest-neighbor approaches feel natural.

In time-series environments both global and local smoothers can be useful for prediction. But there's a twist. Economic time-series data tend to trend, so that x^* can often be exterior to the observed x's. That can create serious issues for local smoothers, as, for example, there may be no nearby "nearest neighbors"! Polynomial and Fourier global smoothers, in contrast, can be readily extrapolated for short-horizon out-of-sample forecasts. They have

issues of their own for long-horizon forecasts, however, as, for example, all polynomials diverge either to $+\infty$ or $-\infty$ when extrapolated far enough.

Nearest Neighbors as a General Forecasting Method

Notice how natural and general NN is for forecasting. If we want to know what y is likely to go with x^* an obvious strategy is to look at the y's that went with x's nearest x^*. And the NN idea can be used to produce not just point forecasts (e.g., by fitting a constant to the y's, but moreover to produce density forecasts (by fitting a distribution to the y's). The NN idea is also equally relevant and useful in time series.

4.4 Wage Prediction, Continued

4.4.1 Point Wage Prediction

4.4.2 Density Wage Prediction

4.5 Exercises, Problems and Complements

1. Additional insight on parameter-estimation uncertainty.

   Consider a simple homogeneous linear regression with zero-mean variables and Gaussian disturbances

$$y_t = \beta x_t + \varepsilon_t$$

$$\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2).$$

   It can be shown that

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2},$$

and that $\hat{\beta}$ and ε are independent. Now consider an operational point prediction of y given that $x = x^*$, $\hat{y} = \hat{\beta}x^*$, and consider the variance of the corresponding forecast error. We have

$$var(e) = var(y - \hat{y}) = var((\beta x^* + \varepsilon) - \hat{\beta}x^*) = \sigma^2 + \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2} x^{*2}.$$

In this expression, the first term accounts for for the usual disturbance uncertainty, and the second accounts for parameter estimation uncertainty. Taken together, the results suggest an operational density forecast that accounts for parameter uncertainty,

$$y_i \,|\, x_i = x^* \;\sim\; N\left(\hat{\beta}x^*,\; \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{\sum_{i=1}^{N} x_i^2}\; x^{*2}\right),$$

from which interval forecasts may be constructed as well. Note that when parameter uncertainty exists, the closer x^* is to the mean x (0), the smaller is the prediction-error variance. Note also that as the sample size gets large, $\sum_{i=1}^{N} x_t^2$ gets large as well, so the adjustment for parameter estimation uncertainty vanishes (in fact very quickly, like $1/N$), and the density forecast collapses to the feasible Gaussian density forecast introduced in the text.

The ideas sketched here can be shown to carry over to more complicated situations (e.g., non-Gaussian, y and x don't necessarily have zero means, more than one regressor, etc.); it remains true that the closer is x to its mean, the tighter is the prediction interval.

2. Prediction intervals via quantile regression.

   Granger, C.W.J., H. White, and M. Kamstra (1987), "Interval Forecasting: An Analysis Based Upon ARCH − Quantile Estimators," Journal of Econometrics. White (1990) allows for nonlinear conditional quantile regression via neural nets.

3. In-sample vs. out-of-sample prediction.

   In cross sections all prediction has an "in-sample" flavor insofar as the X^* for which we want to forecast y is typically interior to the historical X. In time series, in contrast, future times are by definition exterior to past times.

4. Model uncertainty.

   We have thus far emphasized disturbance uncertainty and parameter estimation uncertainty (which is due in part to data uncertainty, which in turn has several components).

   A third source of prediction error is **model uncertainty**. All our models are intentional simplifications, and the fact is that different models produce different forecasts. Despite our best intentions, and our use of powerful tools such as information criteria, we never know the DGP, and surely any model that we use is *not* the DGP.

5. "Data-rich" environments.

   "Big data." "Wide data," for example, corresponds to K large relative to T. In extreme cases we might even have K much larger than T. How to get a sample covariance matrix for the variables in X? How to run a regression? One way or another, we need to recover degrees of freedom, so dimensionality reduction is key, which leads to notions of variable selection and "sparsity", or shrinkage and "regularization".

6. Neural Networks

   Neural networks amount to a particular non-linear functional form associated with repeatedly running linear combinations of inputs through non-linear "squashing" functions. The 0-1 squashing function is useful in classification, and the logistic function is useful for regression. The neural net literature is full of biological jargon, which serves to obfuscate

rather than clarify. We speak, for example, of a "single-output feedforward neural network with n inputs and 1 hidden layer with q neurons." But the idea is simple. If the output is y and the inputs are x's, we write

$$y_t = \Phi(\beta_0 + \sum_{i=1}^{q} \beta_i h_{it}),$$

where

$$h_{it} = \Psi(\gamma_{i0} + \sum_{j=1}^{n} \gamma_{ij} x_{jt}), i = 1, ..., q$$

are the "neurons" ("hidden units"), and the "activation functions" Ψ and Φ are arbitrary, except that Ψ (the squashing function) is generally restricted to be bounded. (Commonly $\Phi(x) = x$.) Assembling it all, we write

$$y_t = \Phi\left(\beta_0 + \sum_{i=1}^{q} \beta_i \Psi\left(\gamma_{i0} + \sum_{j=1}^{n} \gamma_{ij} x_{jt}\right)\right) = f(x_t; \theta),$$

which makes clear that a neural net is just a particular non-linear functional form for a regression model.

7. Trees

8. Kernel Regression

9. Regression Splines

   Polynomial are global. Unattractive in that the fit at the end is influenced by the data at beginning (for example).

   Move to piecewise cubic (say). But it's discontinuous at the join point(s) ("knots").

   Move to continuous piecewise cubic; i.e., force continuity at the knots. But it might have an unreasonable kink.

Move to cubic spline. Forces continuity *and* continuity of first and second derivatives at the knots. Nice! A polynomial of degree p spline has continuous d^{th}-order derivatives, $d = 0, ..., p-1$. So, for example, linear spline is piecewise linear, continuous but not differentiable at the knots.

– Linear Splines

– Constructing Cubic Splines

– Natural Cubic Splines

Extrapolates linearly beyond the left and right boundary knots. This adds constraints (two at each end), recovering degrees of freedom and hence allowing for more knots.

A cubic spline with K knots uses $K + 4$ degrees of freedom. A natural spline with K knots uses K degrees of freedom.

– Knot Placement

You'd like more knots in rough areas of the function being estimated, but of course you don't know where those areas are, so it's tricky.

Smoothing splines avoid that issue.

10. Smoothing Splines

$$\min_{\{f \in F\}} \sum_{t=1}^{T} (y_t - f(x_t))^2 + \lambda \int f''(z)^2 dz$$

HP Trend does that:

$$\min_{\{s_t\}_{t=1}^{T}} \sum_{t=1}^{T} (y_t - s_t)^2 + \lambda \sum_{t=2}^{T-1} \left((s_{t+1} - s_t) - (s_t - s_{t-1})\right)^2$$

The smoothing spline is a natural cubic spline. It has a knot at each unique x value, but smoothness is imposed via λ. No need to choose knot locations; instead just choose a single λ. Can be done by CV.

There is an analytic formula giving effective degrees of freedom, so we can specify d.f. rather than λ.

4.6 Notes

"LOWESS"