22 Lecture 22: March 23

Last time

- Diagnosing non-normality, non-constant error variance (JF chapter 12)

Today

- Diagnosing nonlinearity (JF chapter 12)

Weighted-least-squares estimation

Weighted-least-squares (WLS) regression provides an alternative approach to estimation in the presence of nonconstant error variance. Suppose that the errors from the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ are independent and normally distributed, with zero means but *different* variances: $\epsilon_i \sim N(0, \sigma_i^2)$. Suppose further that the variances of the errors are known up to a constant of proportionality σ_ϵ^2, so that $\sigma_i^2 = \sigma_\epsilon^2 / w_i^2$. Then the likelihood for the model is

$$L(\boldsymbol{\beta}, \sigma_\epsilon^2) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right]$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors,

$$\boldsymbol{\Sigma} = \sigma_\epsilon^2 \times \text{diag}\{1/w_1^2, \ldots, 1/w_n^2\} \equiv \sigma_\epsilon^2 \mathbf{W}^{-1}$$

The maximum-likelihood estimators of $\boldsymbol{\beta}$ and σ_ϵ^2 are then

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y} \\ \hat{\sigma}_\epsilon^2 &= \frac{\sum (w_i \hat{\epsilon}_i)^2}{n}\end{aligned}$$

Correcting OLS standard errors for nonconstant variance

The covariance matrix of the ordinary-least-squares (OLS) estimator is

$$\begin{aligned}\mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Var}\left(\mathbf{Y}\right)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

under the standard assumptions, including the assumption of constant error variance, $\mathbf{Var}\left(\mathbf{Y}\right) = \sigma_\epsilon^2 \mathbf{I}_n$. If, however, the errors are heteroscedastic but independent then $\boldsymbol{\Sigma} \equiv \mathbf{Var}\left(\mathbf{Y}\right) = \text{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, and

$$\mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

White (1980) shows that the following is a consistent estimator of $\mathbf{Var}\left(\hat{\boldsymbol{\beta}}\right)$

$$\tilde{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\Sigma}}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

with $\hat{\mathbf{\Sigma}} = \text{diag}\{\hat{\sigma}_1^2, \ldots, \hat{\sigma}_n^2\}$, where $\hat{\sigma}_i^2$ is the OLS residual for observation i.

Subsequent work suggested small modifications to White's coefficient-variance estimator, and in particular simulation studies by Long and Ervin (2000) support the use of

$$\tilde{\text{Var}}^*(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Sigma}}^*\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

where $\hat{\mathbf{\Sigma}}^* = \text{diag}\{\hat{\sigma}_i^2/(1-h_i)^2\}$ and h_i is the hat-value associated with observation i. In large samples, where h_i is small, the distinction between $\tilde{\text{Var}}(\hat{\boldsymbol{\beta}})$ and $\tilde{\text{Var}}^*(\hat{\boldsymbol{\beta}})$ essentially disappears.

A rough *rule* is that nonconstant error variance seriouly degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more (or, more conservatively, about 4 or more).

Nonlinearity

If $\mathbf{E}(\mathbf{Y}|\mathbf{X})$ is not linear in $\mathbf{X}$ (in other words, $\mathbf{E}(\epsilon|\mathbf{X}) \neq 0$ for some x), $\hat{\boldsymbol{\beta}}$ may be biased and inconsistent. Usually we employ "linearity by default" but we should try to make sure this is appropriate: **detect** non-linearities and **model** them accurately.

Lowess smoother, JF 2.3

We can employ local averaging plots to help with diagnostics. Lowess method is in many respects similar to local-averaging smoothers, except that instead of computing an average Y-value within the neighborhood of a focal x, the lowess smoother computes a *fitted* value based on a locally weighted least-squares line, giving more weight to observations in the neighborhood that are close to the focal x than to those relatively far away. The name "lowess" is an acronym for *lo*cally *we*ighted *s*catterplot *s*moother and is sometimes rendered as *loess*, for *lo*cal regr*ess*ion.

Component-plus-residual plots

Added-variable plots, introduced for detecting influential data, can reveal nonlinearity. However, the added-variable plots are not always useful for locating a transformation:

- The added-variable plot adjusts X_j for the other Xs.
- The *unadjusted* X_j is transformed in respecifying the model.

Moreover, Cook (1998, Section 14.5) shows that added-variable plots are biased toward linearity when the correlations among the explanatory variables are large.

Component-plus-residual plots (also called *partial-residual plots*) are often an effective alternative. The component-plus-residual plots are not as suitable as added-variable plots for revealing leverage and influence, though. The component-plus-residual plots are constructed by

1. Compute residuals from full regression:

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

2. Compute "linear component" of the partial relationship:

$$C_i = \hat{\beta}_j X_{ij}$$

3. Add linear component to residual to get partial residual for the jth explanatory variable

$$\hat{\epsilon}_i^{(j)} = \hat{\epsilon}_i + C_i = \hat{\epsilon}_i + \hat{\beta}_j X_{ij}$$

4. Plot $\hat{\epsilon}_{\cdot}^{(j)}$ against $X_{\cdot j}$

Figure 22.1 shows the component-plus-residual plots for the regression of log wages on variables (age, education and sex) of the 1994 wave of Statistics Canada's Survey of Labour and Income Dynamics (SLID) data. The SLID data set includes 3997 employed individuals who were between 16 and 65 years of age and who resided in Ontario.

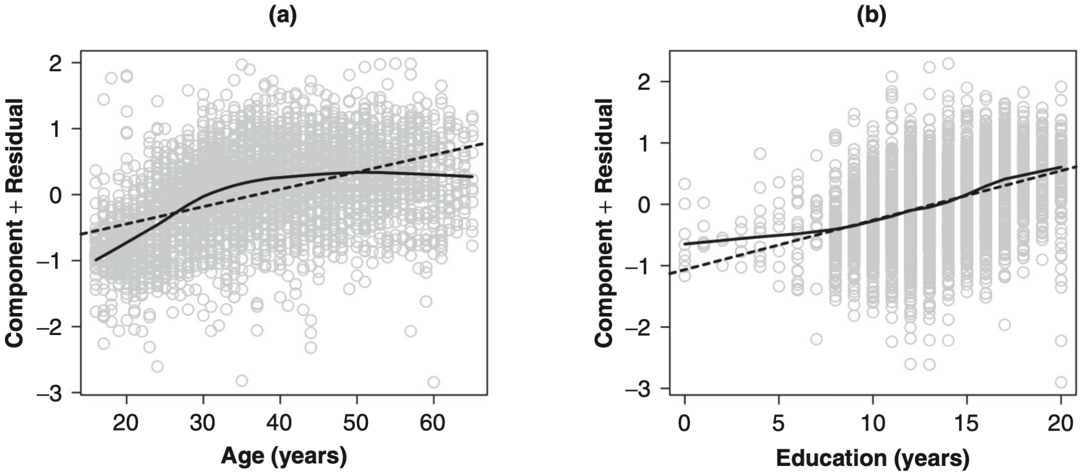

Figure 22.1: Component-plus-residual plots for age and education in SLID regression of log wages on these variables and sex. The solid lines are for lowess smooths with spans of 0.4, and the broken lines are for linear least-squares fits. JF Figure 12.6.

Data transformation

The family of powers and Roots, JF 4.1

A particularly useful group of transformations is the "family" of powers and roots:

$$X \to X^p$$

wehre the arrow indicates that we intend to replace X with the transformed variable X^p. If p is negative, then the transformation is an inverse power. For example, $X^{-1} = 1/X$. If p is a fraction, then the transformation represents a root. For example, $X^{1/3} = \sqrt[3]{X}$.

It is more convenient to define the family of power transformations in a slightly more complex manner, called the Box-Cox family of transformations (introduced in a seminal paper on transformations by Box & Cox, 1964):

$$X \rightarrow X^{(p)} = \frac{X^p - 1}{p}$$

Because $X^{(p)}$ is a linear function of X^p, the two transformations have the same essential effect on the data, but, as is apparent in Figure 22.2

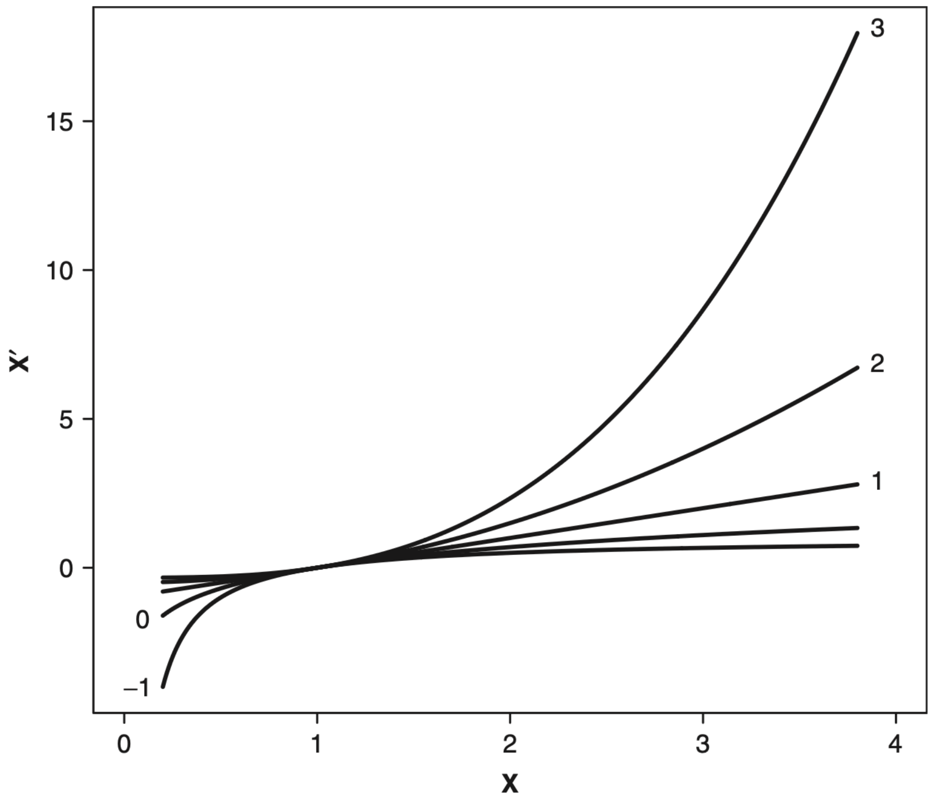

Figure 22.2: The Box-Cox family of power transformations X' of X. The curve labeled p is the transformation $X^{(p)}$, that is $(X^p - 1)/p$; $X^{(0)}$ is $\log_e(X)$. JF Figure 4.1.

- Dividing by p preserves the direction of X, which otherwise would be reversed when p is negative.
- The transformations $X^{(p)}$ are "matched" above $X = 1$ both in level and in slope:
  1. $1^{(p)} = 0$, for all values of p
  2. each transformation has a slope of 1 at $X = 1$.

- Descending the "ladder" of powers and roots towards $X^{(-1)}$ compresses the large values of X and spreads out the small ones. Ascending the ladder of powers and roots towards $X^{(2)}$ has the opposite effect. As p moves further from $p = 1$ (i.e. no transformation) in either direction, the transformation grows more powerful, increasingly "bending" the data.
- The power transformation X^0 is useless because it changes all values to 1, but we can think of the log transformation as a kind of "zeroth" power:

  $$\lim_{p \to 0} \frac{X^p - 1}{p} = \log_e X$$

  and by convention, $X^{(0)} \equiv \log_e X$.

Box-Cox transformation of Y

Box and Cox (1964) suggested a power transformation of Y with the object of normalizing the error distribution, stabilizing the error variance, and straightening the relationship of Y to the Xs. The general Box-Cox model is

$$Y_i^{(\lambda)} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, and

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

Note: in statistics, $\log_e$ is often written as log.

For a particular choice of λ, the conditional maximized log-likelihood (see JF 12.5.1 p.324 footnote 55) is

$$\begin{aligned} \log_e L(\beta_0, \beta_1, \ldots, \beta_p, \sigma_\epsilon^2 | \lambda) &= -\frac{n}{2}(1 + \log_e 2\pi) \\ &\quad - \frac{n}{2} \log_e \hat{\sigma}_\epsilon^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log_e Y_i \end{aligned}$$

where $\hat{\sigma}_\epsilon^2(\lambda) = \sum \hat{\epsilon}_i^2(\lambda)/n$ and where $\hat{\epsilon}_i(\lambda)$ are the residuals from the least-squares regression of $Y^{(\lambda)}$ on Xs. The least-squares coefficients from this regression are the maximum-likelihood estimates of $\boldsymbol{\beta}$s conditional on the values of λ.

A simple procedure for finding the maximum-likelihood estimator $\hat{\lambda}$ is to evaluate the maximized $\log_e L$ (called the profile log-likelihood) for a range of values of λ. To test:$H_0 : \lambda = 1$, calculated the likelihood-ratio statistic

$$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \hat{\lambda})]$$

which is asymptotically distributed as χ_1^2 with one degree of freedom under H_0. A 95% confidence interval for λ includes those values for which

$$\log_e L(\lambda) > \log_e L(\lambda = \hat{\lambda}) - 1.92$$

The number 1.92 comes from $\frac{1}{2}\chi^2_{1,0.05} = 0.5 \times 1.96^2$.

Figure 22.3 shows a plot of the profile log-likelihood against λ for the original SLID regression of composite hourly wages on sex, age, and education. The maximum-likelihood estimate of λ is $\hat{\lambda} = 0.09$, and a 95% confidence interval runs from 0.04 to 0.13. Although 0 is outside of the CI (confidence interval), it is essentially the same transformation of wages as $\lambda = 0.09$ (the correlation between log wages and wages$^{0.09}$ is 0.9996).

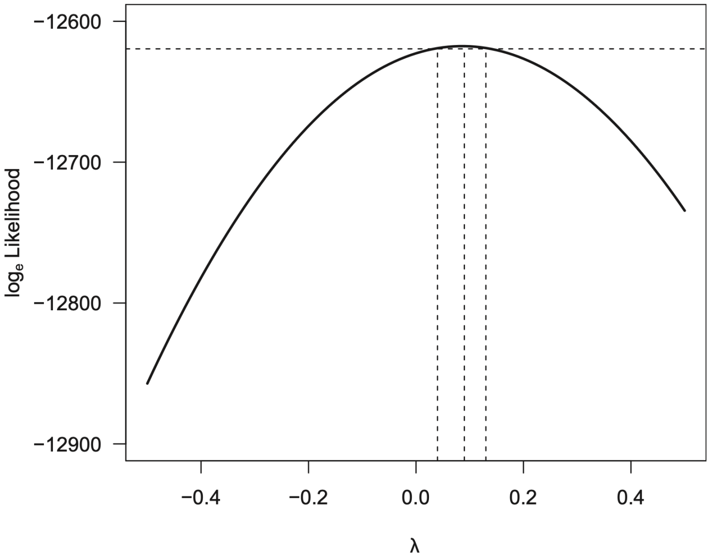

Figure 22.3: Box-Cox transformations for the SLID regression of wages on sex, age, and education. The maximized (profile) log-likelihood is plotted against the transformation parameter λ. The intersection of the line near the top of the graph with the profile log-likelihood curve marks off a 95% confidence interval for λ. The maximum of the log-likelihood corresponds to the MLE of λ. JF Figure 12.14.

Box-Tidwell transformation of Xs

Now, consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_p X_{ip}^{\gamma_p} + \epsilon_i$$

where the errors are independently distributed as $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$ and all the X_{ij} are positive.

The parameters of this model ($\beta_0, \beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p$, and σ^2_ϵ) could be estimated by general nonlinear least squares. Box and Tidwell (1962) suggested the following computationally more efficient procedure (also yields a constructed-variable diagnostic):

1. Regress Y on $X_1, \ldots, X_p$, obtaining $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$. ("Regress A on Bs" is the same as "fitting the linear regression model with A as the response variable and Bs as the explanatory variables".)

2. Regress Y on $X_1, \ldots, X_p$ and the constructed variables $X_1 \log_e X_1, \ldots, X_p \log_e X_p$ (again, by fitting the model of $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \delta_1 X_1 \log_e X_1 + \cdots + \delta_p X_p \log_e X_p + \epsilon_i$)

to obtain $\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_p, \tilde{\delta}_1, \ldots, \tilde{\delta}_p$. In general $\hat{\beta}_i \neq \tilde{\beta}_i$. (The constructed variables result from the first-order Taylor-series approximation to $X_j^{\gamma_j}$ evaluated at $\gamma_j = 1$: $X_j^{\gamma_j} \approx X_1 + (\gamma_1 - 1) X_1 \log_e X_1$.)

3. The constructed variable $X_j \log_e X_j$ can be used to assess the need for a transformation of X_j by testing the null hypothesis $H_0 : \delta_j = 0$. Added-variable plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the Xs.

4. A preliminary estimate of the transformation parameter γ_j (not the MLE) is

$$\tilde{\gamma_j} = 1 + \frac{\tilde{\delta}_j}{\hat{\beta}_j}$$

where $\tilde{\delta}_j$ is from step 2 and $\hat{\beta}_j$ is from step 1.

Polynomial regression

A machinery of multiple regression to fit non-linear relationships between predictor(s) and response.

- Linear: $y = \beta_0 + \beta_1 x + \epsilon$
- Quadratic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- Cubic: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
- k^{th} order polynomial: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$

Question:
Does quadratic model provide a significantly better fit than linear model?
Solution: Test $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$.
Alternatively, compare the corresponding adjusted-R^2 values.