Homework #2

Chapter 2, Shumway and Stoffer

Details of the computing are in the attached R file.

2.1 Johnson & Johnson quarterly earnings

(a) I prefer to use a factor for these sorts of dummy-variable regression models; just a matter of taste. Using the factor, however, is more important when you want to test (partial F test) the cumulative benefit of the collection of indicators.

(b) If you add the intercept, you have a redundant explanatory variable since the sum $Q_1 + Q_2 + Q_3 + Q_4 = 1$. The model is not identified.

(c) The model explains a lot of variation, but leaves an interesting trend in the residuals that suggests that the overall rate of growth represented by $\beta$ in the model has slowed after $t \approx 40$. Seasonal differences are also apparent. Perhaps an interaction is called for? (If so, using a factor in R for the indicators is much easier to introduce these into the model.)

2.4 KL Divergence (a.k.a., relative entropy, log-likelihood ratio)

The divergence measures the “distance” between two distributions (it is not a true distance measure because it is not symmetric; one distribution defines the probability measure). If $f$ is the data generating model, then the divergence from $f$ to another density $g$ is (S&S add a $1/n$ out front)

$$D(f \parallel g) = \mathbb{E}_f \log \frac{f}{g} = \int_x \log \left( \frac{f(x)}{g(x)} \right) f(x) dx$$

(1)

The log of the likelihood ratio, everyone’s favorite test statistic, sits in the integral, so we can think of the divergence as the expected likelihood ratio. (Information theory interprets $D$ as a measure of the inefficiency of coding data using a model based on $g$ when the data were generated by $f$.)

If the data follow a Gaussian regression model $Y = X\beta + \epsilon$, then with parameters $\theta = (\beta, \sigma^2)$ we obtain

$$f_\theta(y) = e^{-\frac{\sum (y_i - x_i'\beta)^2}{2\sigma^2}} \left(\frac{2\pi\sigma^2}{n}\right)^{n/2}$$

(2)

Now simplify after cancelling out common constants noting that $\mathbb{E}_{1}(y_i - x_i'\beta_1)^2 = \sigma_1^2$,

$$\mathbb{E}_1 \log \frac{f_1(y)}{f_2(y)} = \frac{n}{2} \log \frac{\sigma_2^2}{\sigma_1^2} + \mathbb{E}_1 \frac{\sum(y_i - x_i'\beta_2)^2}{2\sigma_2^2} - \mathbb{E}_1 \frac{\sum(y_i - x_i'\beta_1)^2}{2\sigma_1^2}$$

$$= \frac{n}{2} \left( \log \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) + \mathbb{E}_1 \frac{\sum(y_i - x_i'\beta_2)^2}{2\sigma_2^2}$$

$$= \frac{n}{2} \left( \log \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) + \mathbb{E}_1 \frac{\sum(y_i \pm x_i'\beta_1 - x_i'\beta_2)^2}{2\sigma_2^2}$$
\begin{align*}
&= \frac{n}{2} \left( \log \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) + \mathbb{E}_1 \sum (y_i - x'_i \beta_1)^2 + x'_i (\beta_1 - \beta_2)^2 + (y_i - x'_i \beta_1) x'_i (\beta_1 - \beta_2) \frac{2 \sigma_2^2}{\sigma_1^2} \\
&= \frac{n}{2} \left( \log \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) + \sum \sigma_1^2 + x'_i (\beta_1 - \beta_2)^2 \frac{2 \sigma_2^2}{\sigma_1^2} \\
&= \frac{n}{2} \left( \sigma_1^2 - \log \frac{\sigma_2^2}{\sigma_1^2} - 1 \right) + \sum x'_i (\beta_1 - \beta_2)^2 \frac{2 \sigma_1^2}{\sigma_2^2} \\
\end{align*}

With $X$ in place of $Z$ and summations in place of dot products, this matches expression (2.56).

### 2.5 AIC model selection

This problem considers using the divergence $D(f_\theta(y) \| f_{\hat{\theta}(y)}(y))$ as a means to model selection. (To get a sense that all is not well with this approach, notice that $f_{\hat{\theta}(y)}(y)$ is not a density function (it integrates to more than 1; for example, integrate the normal density with $\sigma^2 = 1$ and $\mu = \overline{y}$.)

The expected value sought is (using the result of 2.4)

\[ E_1 \log \frac{f_\theta(y)}{f_{\hat{\theta}}(y)} = \frac{n}{2} \left( \log \frac{\sigma^2}{\sigma_2^2} - \log \frac{\sigma_1^2}{\sigma_2^2} - 1 \right) + \frac{k \sigma_2^2}{2 \sigma_1^2} \]  

(3)

The problem gives the key hint needed for finding the expected value of the reciprocal of a chi-squared r.v. Using this, we get

\[ \mathbb{E} \frac{\sigma^2}{\sigma_2^2} = n \mathbb{E} \frac{\sigma^2}{n \sigma_2^2} = \frac{n}{n - k - 2}. \]

Using the second hint and independence of $\hat{\sigma}$ and $\hat{\beta}$,

\[ \mathbb{E} \sum \frac{x'_i (\beta - \hat{\beta})^2}{\sigma^2} \frac{\sigma_1^2}{2 \sigma_2^2} = k \frac{n}{2(n - k - 2)} \]

Plug these back into (3) and you get

\[ E_1 \log \frac{f_\theta(y)}{f_{\hat{\theta}}(y)} = \frac{n}{2} \left( \frac{n}{n - k - 2} - \mathbb{E} \log \frac{\sigma^2}{\sigma_2^2} - 1 + \frac{k}{n - k - 2} \right) \]

\[ = \frac{n}{2} \left( \mathbb{E} \log \frac{\sigma_1^2}{\sigma_2^2} - 1 + \frac{n + k}{n - k - 2} \right) \]

The last term produces the familiar “two times the number of parameters” penalty since $(n + k)/(n - k - 2) \approx (n + k)/(n - k) = 1 + 2k/(n - k)$.

### 2.9 Oil and gas prices

(a) The timeplots show a very slowly meandering pattern, with periods of long trends. The slow decay of the ACFs suggests long-term dependence in the prices. (Finance would argue for a random walk.)

(b) Returns are a good thing. Notice that the percentage change is almost the same as the log of the ratio. Write $x_{t+1} = \delta_t + x_t$ and assume $\delta_t \ll x_{t-1}$; then

\[ \log \frac{x_t}{x_{t-1}} = \log \frac{x_{t-1} + \delta_t}{x_{t-1}} \]
\[ = \log(1 + \frac{\delta_t}{x_{t-1}}) \]
\[ \approx \frac{\delta_t}{x_{t-1}} \]

which is the percentage change after we multiply by 100.

(c) Plots of the cross-correlations are relatively symmetric, suggesting neither series leads or lags
the other by a noticeable amount. The aligned timeplots of the returns suggest the only
interesting places to look are those with large changes. (Oil returns are almost constant
during this time period.)

2.12 Global temperature

Various smoothers seem to find a kink or elbow near 1910. Some smoothers provide a rougher
fit than others (with more wiggles here and there), but the major elbow is consistently positioned
here.