Emily
de Vibe

MATH581

Coursework 2

1. The volatility of a stock or group of stocks can be estimated from the price of options for that stock, or group of stocks. This involves, for example, inverting the Black-Scholes formula for pricing options to obtain the value of the volatility that produces the given price for the option. If we are interested in the volatility of (say) the S&P500, we can do this for multiple stocks in the S&P500 and average them. These estimates are called *implied volatility*.

The VIX and VXO indices show estimates of the implied volatility for the S&P500

Consider data on log-returns of a share in the S&P500. We wish to model the variability of these returns as a function of the VIX index. Let $y_1, \dots, y_n$ denote the observed daily log-returns, and $z_1, \dots, z_n$ the value of the VIX index at the start of each daily period. Our model is

$$y_t \sim N(\mu, \alpha(1 + \beta z_t))$$

(a) Write down the log-likelihood for the data under this model

Our model is

$$y_t \sim N(\mu, \alpha(1+\beta z_t))$$

so that $\vec{\theta} = (\mu, \alpha)$ (assuming β is known).

Then the likelihood of data $\vec{y} = (y_1, \dots, y_n)$ is

$$L(\vec{\theta}) = \prod_{t=1}^{n} f(x_t | \vec{\theta})$$

$$= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi} \cdot \sqrt{\alpha(1+\beta z_t)}} e^{-\frac{1}{2\alpha(1+\beta z_t)}(y_t - \mu)^2}$$

Therefore the log-likelihood for the data under this model is:

$$\ell(\vec{\theta}) = \sum_{t=1}^{n} \ln(f(x_t | \vec{\theta}))$$

$$= \sum_{t=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}\sqrt{\alpha(1+\beta z_t)}} e^{-\frac{1}{2\alpha(1+\beta z_t)}(y_t - \mu)^2} \right)$$

$$= \sum_{t=1}^{n} \left(\ln(1) - \ln\left((2\pi)^{1/2} (\alpha(1+\beta z_t))^{1/2}\right) - \frac{1}{2\alpha(1+\beta z_t)} (y_t - \mu)^2 \right)$$

$$= \sum_{t=1}^{n} \left(0 - \left(\ln\left((2\pi)^{1/2}\right) + \ln\left((\alpha(1+\beta z_t))^{1/2}\right) \right) - \frac{1}{2\alpha(1+\beta z_t)} (y_t - \mu)^2 \right)$$

$$= \sum_{t=1}^{n} \left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\alpha(1+\beta z_t)) - \frac{1}{2\alpha(1+\beta z_t)} (y_t - \mu)^2 \right)$$

$$= -\frac{n}{2}\ln(2\pi) - \sum_{t=1}^{n} \frac{1}{2}\ln(\alpha(1+\beta z_t)) - \sum_{t=1}^{n} \frac{(y_t - \mu)^2}{2\alpha(1+\beta z_t)}$$

(b) Assume that β is known. Calculate the MLEs for μ and α.

We assume β is known. So calculate the MLEs for μ and α, we differentiate the log-likelihood function $\ell(\vec{\theta})$ with respect to μ and α as following:

$$\frac{\partial \ell(\vec{\theta})}{\partial \mu} = -\sum_{t=1}^{n} \frac{\alpha(y_t - \mu)(-1)}{\alpha(1+\beta z_t)}$$

$$= \sum_{t=1}^{n} \frac{(y_t - \mu)}{\alpha(1+\beta z_t)}$$

$$= \sum_{t=1}^{n} \frac{y_t}{\alpha(1+\beta z_t)} - \sum_{t=1}^{n} \frac{\mu}{\alpha(1+\beta z_t)}$$

$$= \sum_{t=1}^{n} \frac{y_t}{\alpha(1+\beta z_t)} - \mu \cdot \sum_{t=1}^{n} \frac{1}{\alpha(1+\beta z_t)}$$

and

$$\frac{\partial \ell(\vec{\theta})}{\partial \alpha} = -\sum_{t=1}^{n} \frac{1}{2\alpha(1+\beta z_t)} (1+\beta z_t)$$

$$-\sum_{t=1}^{n} \frac{(y_t - \mu)^2}{2(1+\beta z_t)} (-1)\alpha^{-2}$$

$$= -\frac{1}{2\alpha} + \sum_{t=1}^{n} \frac{(y_t - \mu)^2}{2\alpha^2(1+\beta z_t)}$$

$$= -\frac{1}{2\alpha} + \frac{1}{2\alpha^2} \sum_{t=1}^{n} \frac{(y_t - \mu)^2}{(1+\beta z_t)}$$

Since $\theta = \hat{\theta}$ maximizes the likelihood function $\ell(\theta)$, we know that

$$\frac{\partial \ell(\hat{\theta})}{\partial \mu} = 0 = \sum_{t=1}^{n} \frac{y_t}{\hat{\alpha}(1+\beta z_t)} - \hat{\mu} \cdot \sum_{t=1}^{n} \frac{1}{\hat{\alpha}(1+\beta z_t)}$$

and

$$\frac{\partial \ell(\hat{\theta})}{\partial \alpha} = -\frac{1}{2\hat{\alpha}} + \frac{1}{2\hat{\alpha}^2} \sum_{t=1}^{n} \frac{(y_t - \hat{\mu})^2}{(1+\beta z_t)}$$

Solving with respect to $\hat{\mu}$ and $\hat{\alpha}$ yields the following MLEs for μ and α:

$$\hat{\mu} = \bar{y} \quad \text{and} \quad \hat{\alpha} = \sum_{t=1}^{n} \frac{(y_t - \bar{y})^2}{(1+\beta z_t)}$$

Calculamos:

$$\frac{\partial \ell(\hat{\theta})}{\partial \mu} = 0 = \sum_{t=1}^{n} \frac{y_t}{\hat{\alpha}(1+\beta z_t)}$$

$$\hat{\mu} \sum_{t=1}^{n} \frac{1}{\hat{\alpha}(1+\beta z_t)}$$

$$\Rightarrow \hat{\mu} \sum_{t=1}^{n} \frac{1}{\hat{\alpha}(1+\beta z_t)} = \sum_{t=1}^{n} \frac{y_t}{\hat{\alpha}(1+\beta z_t)}$$

$$\Rightarrow \hat{\mu} \cdot n = \sum_{t=1}^{n} y_t$$

$$\Rightarrow \hat{\mu} = \bar{y}$$

$$\frac{\partial \ell(\hat{\theta})}{\partial \beta} = -\frac{1}{\hat{\alpha}} + \frac{1}{\hat{\alpha}^2} \sum_{t=1}^{n} \frac{(y_t - \hat{\mu})^2}{(1+\beta z_t)} = 0$$

$$\Rightarrow \frac{1}{\hat{\alpha}} = \frac{1}{\hat{\alpha}^2} \sum_{t=1}^{n} \frac{(y_t - \hat{\mu})^2}{(1+\beta z_t)}$$

$$\Rightarrow 1 = \frac{1}{\hat{\alpha}} \sum_{t=1}^{n} \frac{(y_t - \hat{\mu})^2}{(1+\beta z_t)}$$

$$\Rightarrow \hat{\alpha} = \sum_{t=1}^{n} \frac{(y_t - \hat{\mu})^2}{(1+\beta z_t)}$$

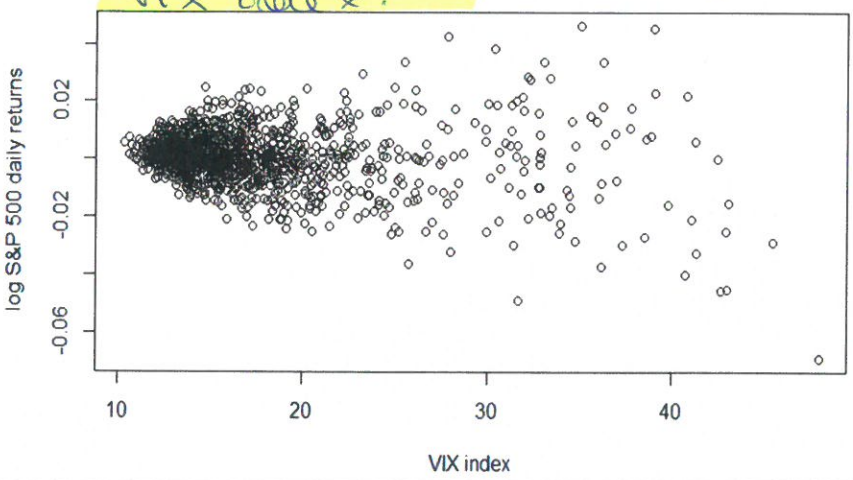
$$\Rightarrow \hat{\alpha} = \frac{\sum_{t=1}^{\hat{}} (y_t - \hat{\mu})^2}{t=1 (1 + \beta z_t)} = \frac{\sum_{t=1}^{\hat{}} (y_t - \bar{y})^2}{(1 + \beta z_t)}$$

(c) In file "SPdata.txt" are the log S&P 500 daily returns (column 1) and VIX index at the start of each daily period (column 2) for 1258 days of trading between 01.07.2011 and 30.06.2016

(i) Plot the log S&P 500 daily returns against VIX index. Does the proposed model look reasonable.

Below follows a plot of the log S&P 500 daily returns against VIX index (Figure 1):

Fig 1 the log S&P 500 daily returns against the VIX index.



our proposed model assumes that $y_+ \sim N(\mu, \alpha(1+\beta z_+))$. By observing Fig. 1, we observe that the mean of y_+ seems located at $\mu = 0$ and remains fixed as the Vix index (at the start of each daily period) increases. Furthermore, the distribution of y_+ seems to be symmetric at $\mu \approx 0$. On the other hand, the width of the distribution of y_+ seems to increase linearly with the Vix index. Therefore it would be reasonable to conclude that $y_+ \sim N(\mu, \alpha(1+\beta z_+))$

(ii) Calculate the MLEs for μ and α and the log likelihood for the MLEs for

$\beta = 0, 1, 10, 100$.

The table below (Table 1) shows the values of the MLEs for μ and α and the log likelihood for the MLEs when

$\beta = 0, 1, 10, 100$

β	0	1	10	100
$\hat{\mu}$	0.0003682654	0.0003682654	0.0003682654	0.0003682654
$\hat{\alpha}$	$1.244101 \cdot 10^{-1}$	$5.443245 \cdot 10^{-3}$	$5.697105 \cdot 10^{-4}$	$5.723936 \cdot 10^{-5}$
$\ell(\hat{\mu}, \hat{\alpha})$	154.4196	315.4186	320.7240	321.2754
$(\hat{\alpha} \cdot \hat{\beta})$	0	0.005443245	0.005697105	0.005723936

Table 1: Calculation results.

(iii) Comment briefly on your results.

We observe that as β increases, the value of $\hat{\mu}$, which is the sample mean of the log S&P 500 daily returns, remains constant since it is only dependent on the values of our data $(y_1, \ldots, y_n)$. However, the value of $\hat{\alpha}$ decreases while the log likelihood for the MLEs increases when the value of β increases.

Furthermore the value of $\hat{\alpha} \cdot \hat{\beta}$ is positive and increases as β increases and therefore the slope of $\sigma^2_{y_t} = \alpha + \alpha\beta z_t$ will also increase when β increases.

2. We let y_i denote the log-wage of individual i and let x_{i1}, x_{i2}, x_{i3}, x_{i4} and x_{i5} denote the corresponding covariates for educ, exper, lnexper, lneduc and male.

(a) Write down the full linear model for y_i.

we fit the model

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \sigma \varepsilon_i$$

we assume ε_i are i.id standard normal random variables.

Thus model is equivalent to assuming

$$y_i \sim N(\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}, \sigma^2)$$

our full linear model can be written in terms of vectors and matrices as following:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \vec{y} = \vec{x}\,\vec{\beta} + \sigma\vec{\varepsilon}$$

$$= \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \sigma \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

we split the data into 2 sets. use the first 1000 observations for fitting the model and the

remaining 472 observations for predictions (test data data set)

(b) Use backwards step-wise selection to choose a reduced linear model for y, using the first 1000 observations, outlining the steps taken. Report coefficient estimates and standard error of parameters.

The coefficien estimates are:

$\hat{\alpha} \approx 4.8506$ with SE ≈ 0.0514,

$\hat{\beta}_1 \approx 0.1519$ with SE ≈ 0.0098,

$\hat{\beta}_3 \approx 0.2144$ with SE ≈ 0.0125,

and $\hat{\beta}_5 \approx 0.1316$ with SE ≈ 0.0179.

(c) By studying the observed (y_i) against the expected ($\hat{y}_i$) log wages and the residuals, comment upon the appropriateness of the model.

By studying the observed (y_i) against the expected ($\hat{y}_i$) log wages in Fig 1 below, we observe that on one side there are some differences between the observed and expected log wages. For instance, if y_i is below the red line $y = x$, then $\hat{y}_i > y_i$, but if y_i is above the line $y = x$, then $y_i > \hat{y}_i$. If all $y_i = \hat{y}_i$, then all y_i in our plot would be located on the line $y = x$. However all y_i has clearly clustered around the line $y = x$ instead of being spread randomly or forming a nonlinear relation.

Fig 1: Plot of observed against expected log wage

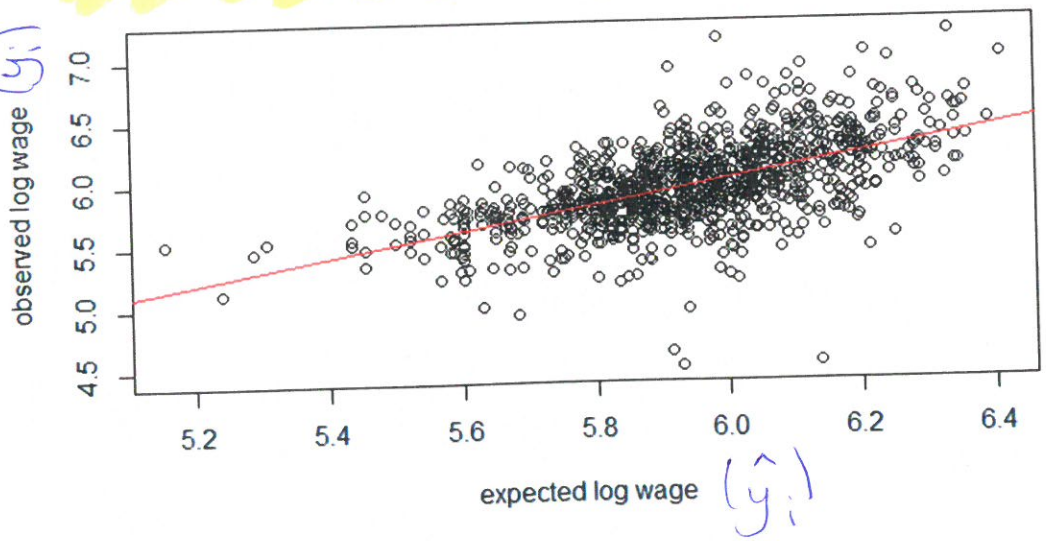
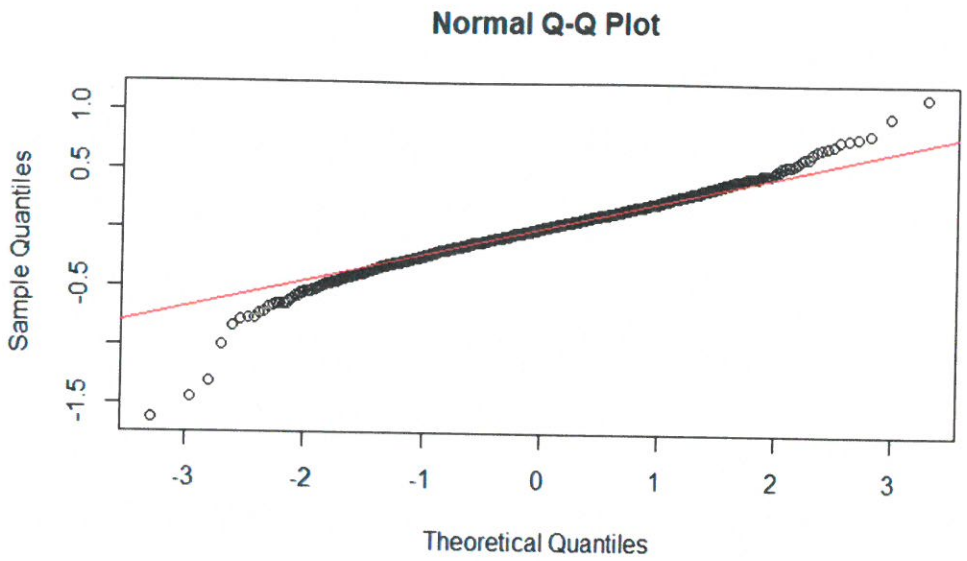


Fig. 2 Normal qqplot of residuals.

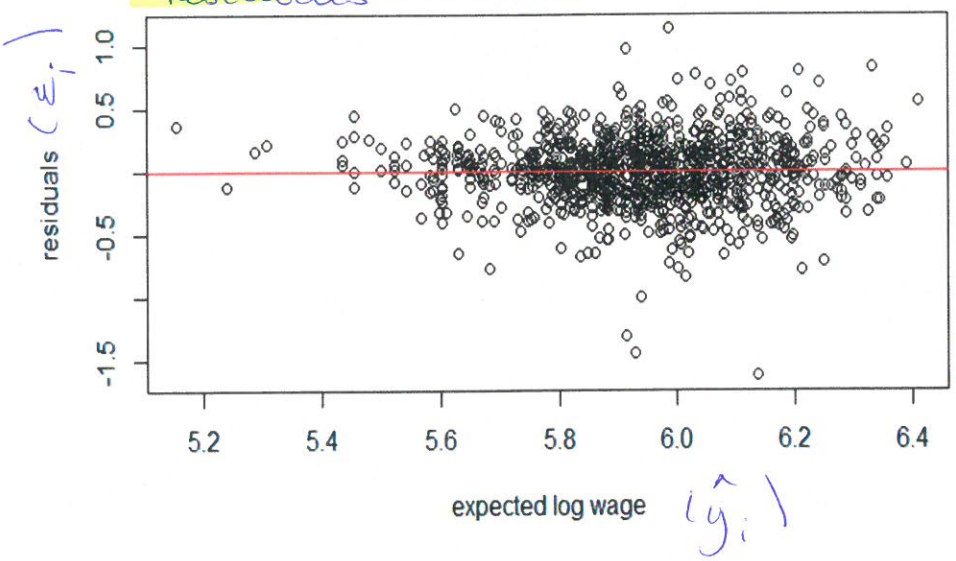


By plotting a normal qqplot of the residuals defined as $\varepsilon_i = y_i - \hat{y}_i$ and as illustrated in Fig. 2, we observe that the quantiles of our residuals ε_i follow the line $y = x$ except for the quantiles of ε_i located on the far left and right extremes of our plot. This means that the quantiles of ε_i except for its lower and upper quantiles are identical to the quantiles of the normal distribution so that the corresponding residuals are normally distributed. ($\varepsilon_i \sim N(0, 1)$) On the other hand, the lower quantiles of the residuals are located below the line $y = x$ indicating that the normal distribution has greater lower quantiles than the quantiles of our residuals. This is because the left tail of the distribution of the normal distribution

is figure than the left tail of the distribution of our residuals. Furthermore, since the upper quantiles of our residuals are located above the line $y = x$, this indicates that the upper quantiles of our residuals are greater than the upper normal quantiles and hence the normal distribution has a heavier right tail than the right tail of the distribution of our residuals. However, the majority of the quantiles of our residuals are located on the line $y = x$ and we must also be aware that all samples can contain outliers which can alter the qq plot of the quantiles of our residuals.

Therefore based on the plot of observed against the expected log wage and the qqplot of our residuals, our model seems appropriate.

Fig. 8: the expected log wage against the residuals



But

if we observe the plot of the expected log wage ($\hat{y}_i$) against the residuals ($\varepsilon_i = y_i - \hat{y}_i$) (Fig 3.) (4), we can see that our residuals are clearly clustered around the red line $y = 0$. However, the residuals should be normally distributed around 0 and therefore they should not show a specific pattern in our plot. As a result, the plot of our residuals against the expected log wages casts doubt on the validity of our model.

As a conclusion, even though the plot of our residuals against the expected log wage argues against the appropriateness of our model, the plot of the observed against the expected log wage and the normal qqplot of our residuals argue for the appropriateness of our model. As different models have both strengths and weaknesses and emphasizes different aspects, our model seems appropriate, but it is not ideal and could be better.

(4) According to Jay L. Devore and Kenneth N. Berk, Modern Mathematical Statistics with Applications (2nd edition), p. 676

Diagnostic Plots

The basic plots that many statisticians recommend for an assessment of model validity and usefulness are the following:

1. y_i on the vertical axis versus x_i on the horizontal axis
2. y_i on the vertical axis versus $\hat{y}_i$ on the horizontal axis
3. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis
4. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis
5. A normal probability plot of the standardized residuals (or residuals)

Plots 3 and 4 are called **residual plots** against the independent variable and fitted (predicted) values, respectively.

If Plot 2 yields points close to the 45° line [slope +1 through (0, 0)], then the estimated regression function gives accurate predictions of the values actually observed. Thus Plot 2 provides a visual assessment of model effectiveness in making predictions. Provided that the model is correct, neither residual plot should exhibit distinct patterns. The residuals should be randomly distributed about 0 according to a normal distribution, so all but a very few standardized residuals should lie between −2 and +2 (i.e., all but a few residuals within 2 standard deviations of their expected value 0). The plot of standardized residuals versus $\hat{y}$ is really a combination of the two other plots, showing implicitly both how residuals vary with x and how fitted values compare with observed values. This latter plot is the single one most often recommended for multiple regression analysis. Plot 5 allows the analyst to assess the plausibility of the assumption that ε has a normal distribution.

Example 12.24 (Example 12.23 continued)

Figure 12.27 presents the five plots just recommended along with a sixth plot. The plot of y versus $\hat{y}$ confirms the impression given by r^2 that x is fairly effective in predicting y. The residual plots show no unusual pattern or discrepant values. The normal probability plot of the standardized residuals is quite straight. In summary, the first five plots leave us with no qualms about either the appropriateness of a simple linear relationship or the fit to the given data.

Notice that plotting against x yields the same shape as a plot against the predicted values. Is this surprising? The predicted value is a linear function of x, so the plots will have the same appearance. Given that the plots look the same, why include both? This is preparation for the next section, where more than one predictor is allowed, and plotting against x is not the same as plotting against the predicted values.

The sixth plot in Figure 12.27 is in accord with what was found graphically in Example 12.12. In that example, Figure 12.18 showed that private universities might tend to have better graduation rates than state universities. For another graphical view of this, we show in the last plot of Figure 12.27 the standardized residuals plotted against a variable that is 0 for state universities and 1 for private universities. In this graph the private universities do seem to have an advantage, but we will need to wait until the next section for a hypothesis test, which requires including this new variable as a second predictor in the model.

d) Generate predictions for the wages of the 472 individuals in the test data (R code is sufficient you do not need to report the predictions). Compute the mean square error and the correlation between the observed and the predicted wages.

By using R (code related to task 2d) we can generate predictions for the wages of the 472 individuals in the test data.

The mean square error is defined as

$$\sigma^2 = \sum_{i=1}^{n} \frac{\varepsilon_i^2}{n}, \text{ where } \varepsilon_i = y - \hat{y}_i \text{ are}$$

the residuals of the fitted model.

Again using R, the mean square error is:

$MSE = \sigma^2 \approx 0.098\,622\,64$

while the correlation between the observed and the predicted wages is $R = 0.556\,007\,9 > 0$ which is a high value indicating a strong positive correlation between the observed and predicted values.