Lecture 7
Linear Regression Diagnostics

BIOST 515

January 27, 2004

Major assumptions

1. The relationship between the outcomes and the predictors is (approximately) linear.

2. The error term ϵ has zero mean.

3. The error term ϵ has constant variance.

4. The errors are uncorrelated.

5. The errors are normally distributed or we have an adequate sample size to rely on large sample theory.

We should always check fitted models to make sure that these assumptions have not been violated.

Departures from the underlying assumptions cannot be detected using any of the summary statistics we've examined so far such as the t or F statistics or R^2. In fact, tests based on these statistics may lead to incorrect inference since they are based on many of the assumptions above.

Residual analysis

The diagnostic methods we'll be exploring are based primarily on the residuals. Recall, the residual is defined as

$$e_i = y_i - \hat{y}_i, \ i = 1, \ldots, n,$$

where

$$\hat{y} = X\hat{\beta}.$$

If the model is appropriate, it is reasonable to expect the residuals to exhibit properties that agree with the stated assumptions.

Characteristics of residuals

- The mean of the $\{e_i\}$ is 0:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0.$$

- The estimate of the population variance computed from the sample of the n residuals is

$$S^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2$$

  which is the residual mean square, $MSE = SSE/(n-p-1)$.

- The $\{e_i\}$ are not independent random variables. In general, if the number of residuals (n) is large relative to the number of independent variables (p), the dependency can be ignored for all practical purposes in an analysis of residuals.

Methods for standardizing residuals

- Standardized residuals
- Studentized residuals
- Jackknife residuals

Standardized residuals

An obvious choice for scaling residuals is to divide them by their estimated standard error. The quantity

$$z_i = \frac{e_i}{\sqrt{MSE}}$$

is called a **standardized residual**. Based on the linear regression assumptions, we might expect the z_is to resemble a sample from a $N(0, 1)$ distribution.

Studentized residuals

Using MSE as the variance of the ith residual e_i is only an approximation. We can improve the residual scaling by dividing e_i by the standard deviation of the ith residual. We can show that the covariance matrix of the residuals is

$$\text{var}(e) = \sigma^2(I - H).$$

Recall $H = X(X'X)^{-1}X'$ is the hat matrix. The variance of the ith residual is

$$\text{var}(e_i) = \sigma^2(1 - h_i),$$

where h_i is the ith element on the diagonal of the hat matrix and $0 \leq h_i \leq 1$.

The quantity

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_i)}}$$

is called a **studentized residual** and approximately follows a t distribution with $n-p-1$ degrees of freedom (assuming the assumptions stated at the beginning of lecture are satisfied).

Studentized residuals have a mean near 0 and a variance,

$$\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2,$$

that is slightly larger than 1. In large data sets, the standardized and studentized residuals should not differ dramatically.

Jackknife residuals

The quantity

$$r_{(-i)} = r_i \sqrt{\frac{MSE}{MSE_{(-i)}}} = \frac{e_i}{\sqrt{MSE_{(-i)}(1-h_i)}} = r_i \sqrt{\frac{(n-p-1)-1}{(n-p-1)-r_i^2}}$$

is called a **jackknife residual** (or **R-Student residual**). $MSE_{(-i)}$ is the residual variance computed with the ith observation deleted.

Jackknife residuals have a mean near 0 and a variance

$$\frac{1}{(n-p-1)-1} \sum_{i=1}^{n} r_{(-i)}^2$$

that is slightly greater than 1. Jackknife residuals are usually the preferred residual for regression diagnostics.

How to use residuals for diagnostics?

Residual analysis is usually done graphically. We may look at

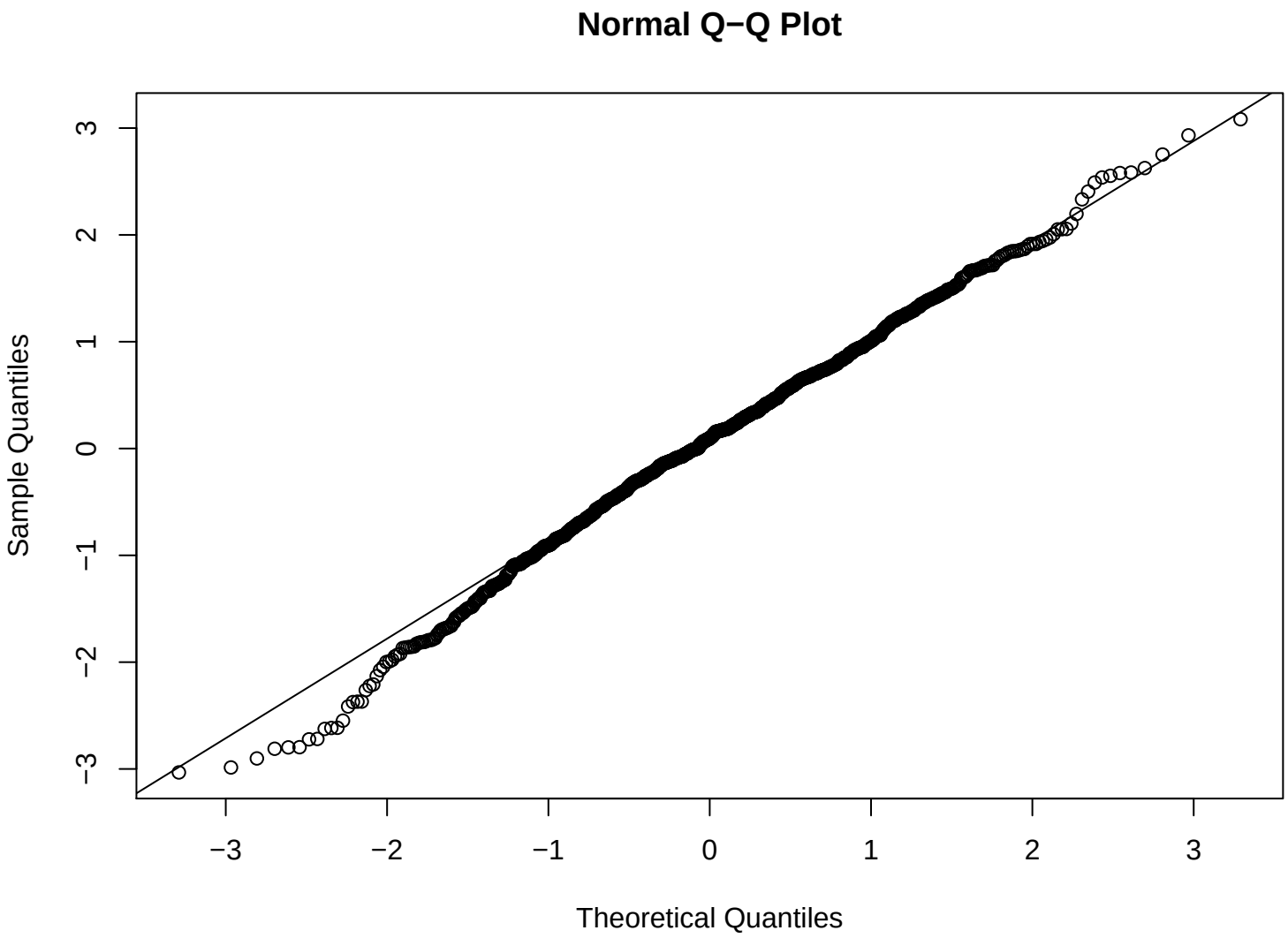
- Quantile plots: to assess normality
- Scatterplots: to assess model assumptions, such as constant variance and linearity, and to identify potential outliers
- Histograms, stem and leaf diagrams and boxplots

Quantile-quantile plots

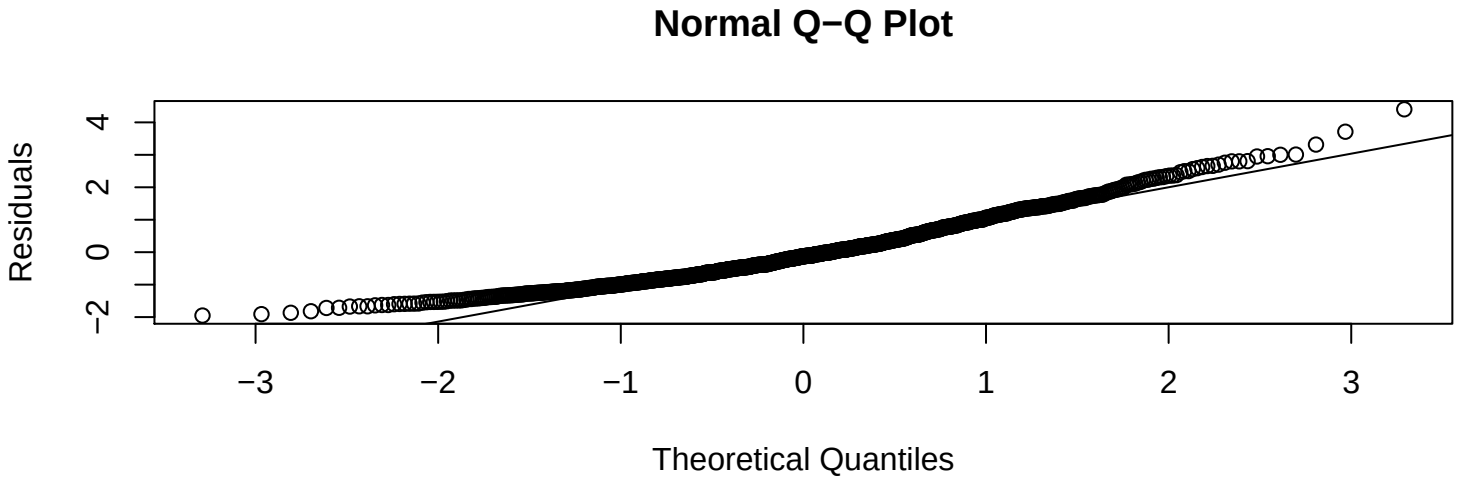
Quantile-quantile plots can be useful for comparing two samples to determine if they arise from the same distribution. Similarly, we can compare quantiles of a sample to the expected quantiles if the sample came from some distribution F for a visual assessment of whether the sample arises from F. In linear regression, this can help us determine the normality of the residuals (if we have relied on an assumption of normality).

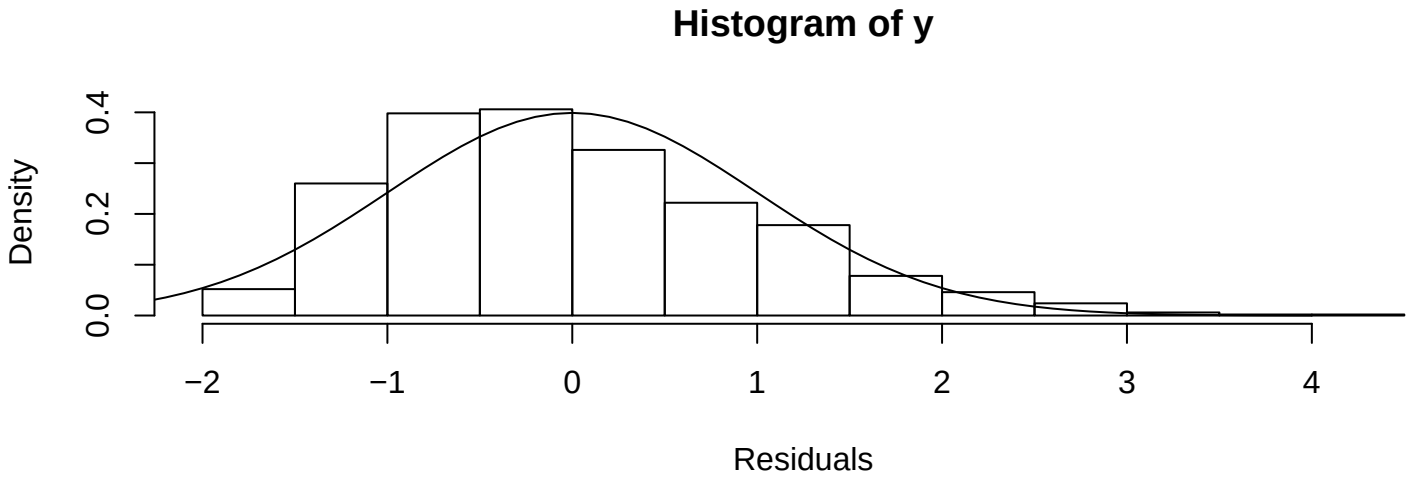
To construct a quantile-quantile plot for the residuals, we plot the quantiles of the residuals against the theorized quantiles if the residuals arose from a normal distribution. If the residuals come from a normal distribution the plot should resemble a straight line. A straight line connecting the 1st and 3rd quartiles is often added to the plot to aid in visual assessment.

Samples from $N(0, 1)$ distribution

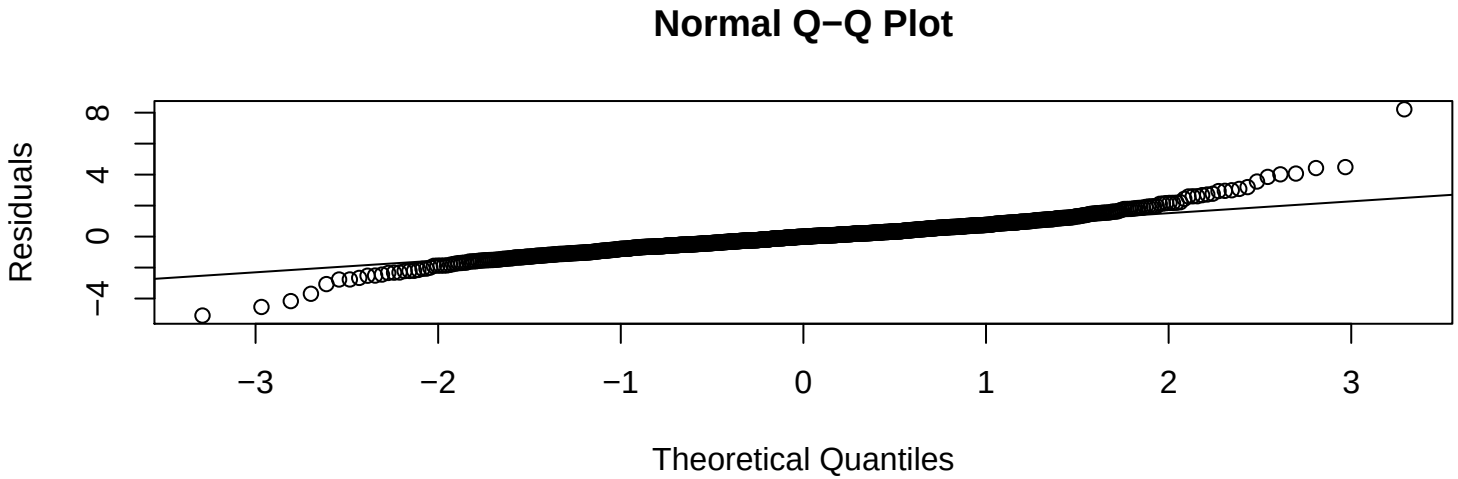


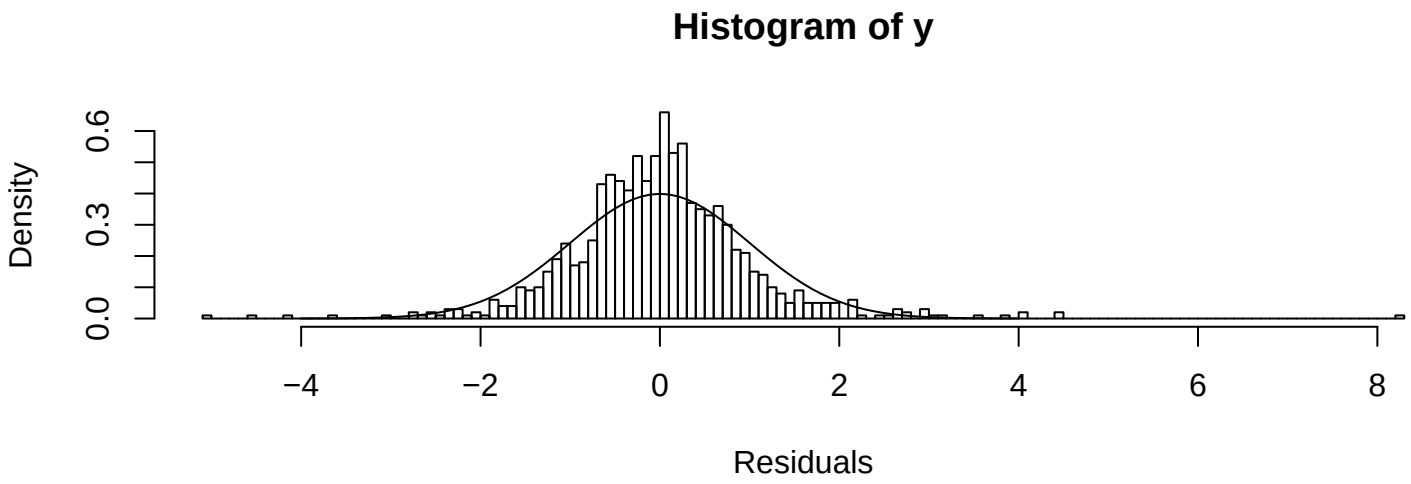
Samples from a skewed distribution



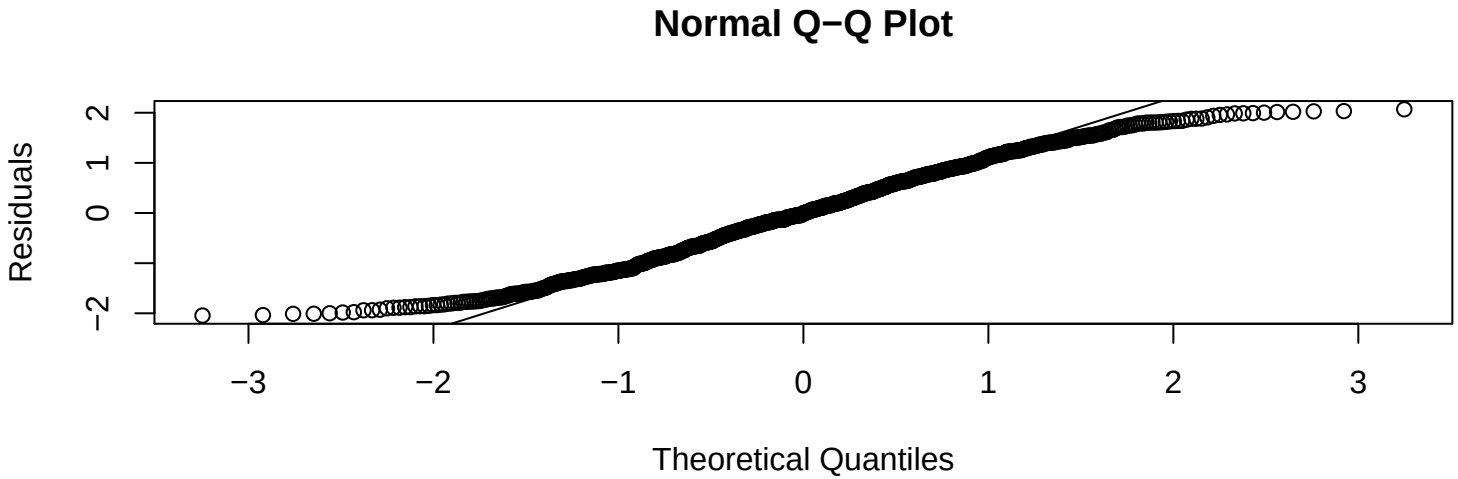


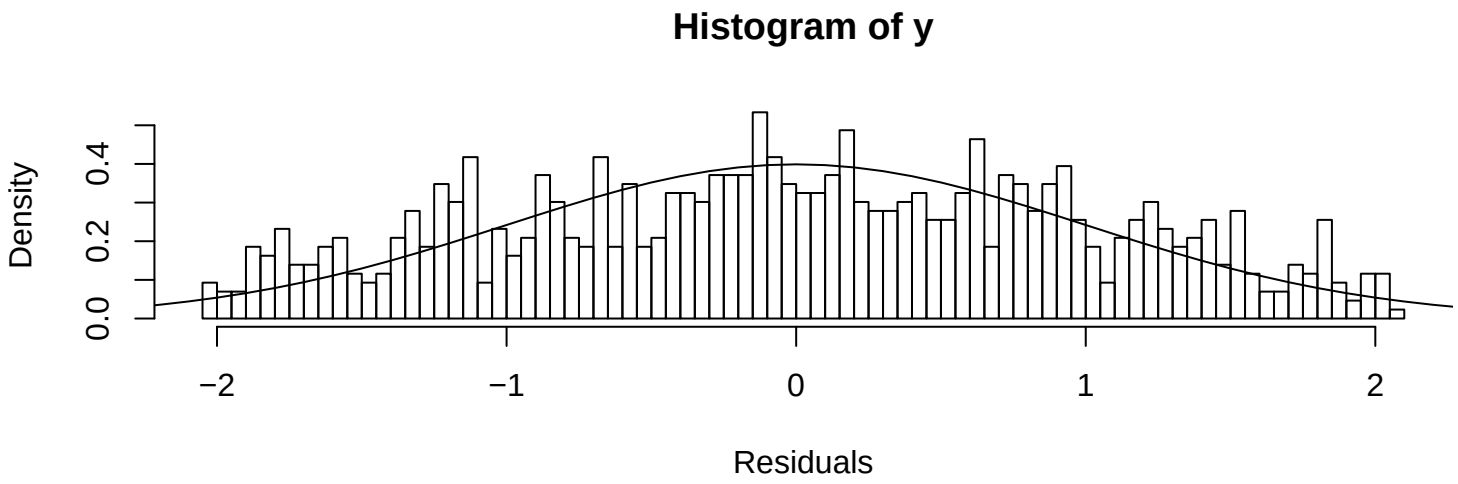
Samples from a heavy-tailed distribution





Samples from a light-tailed distribution



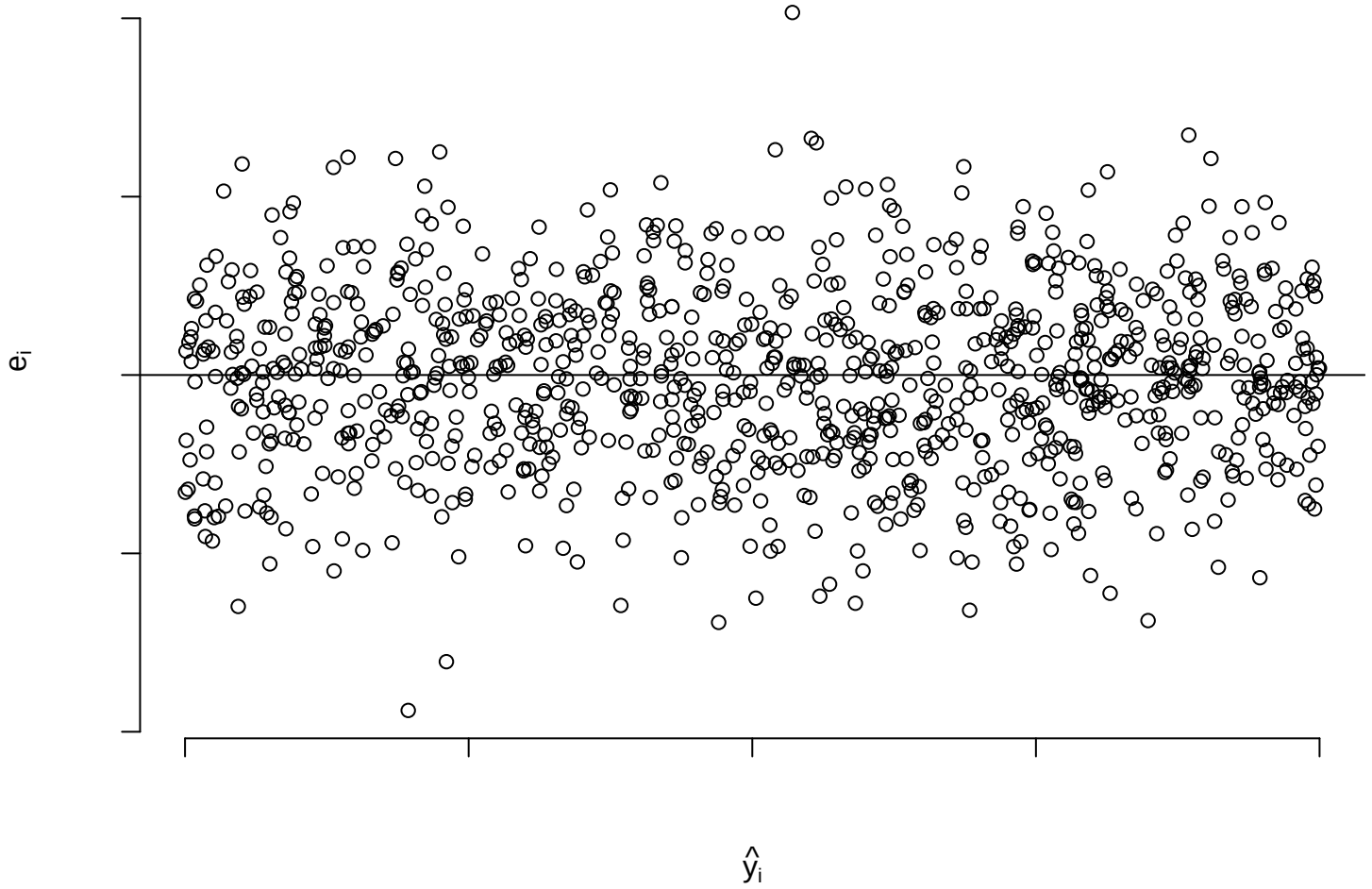


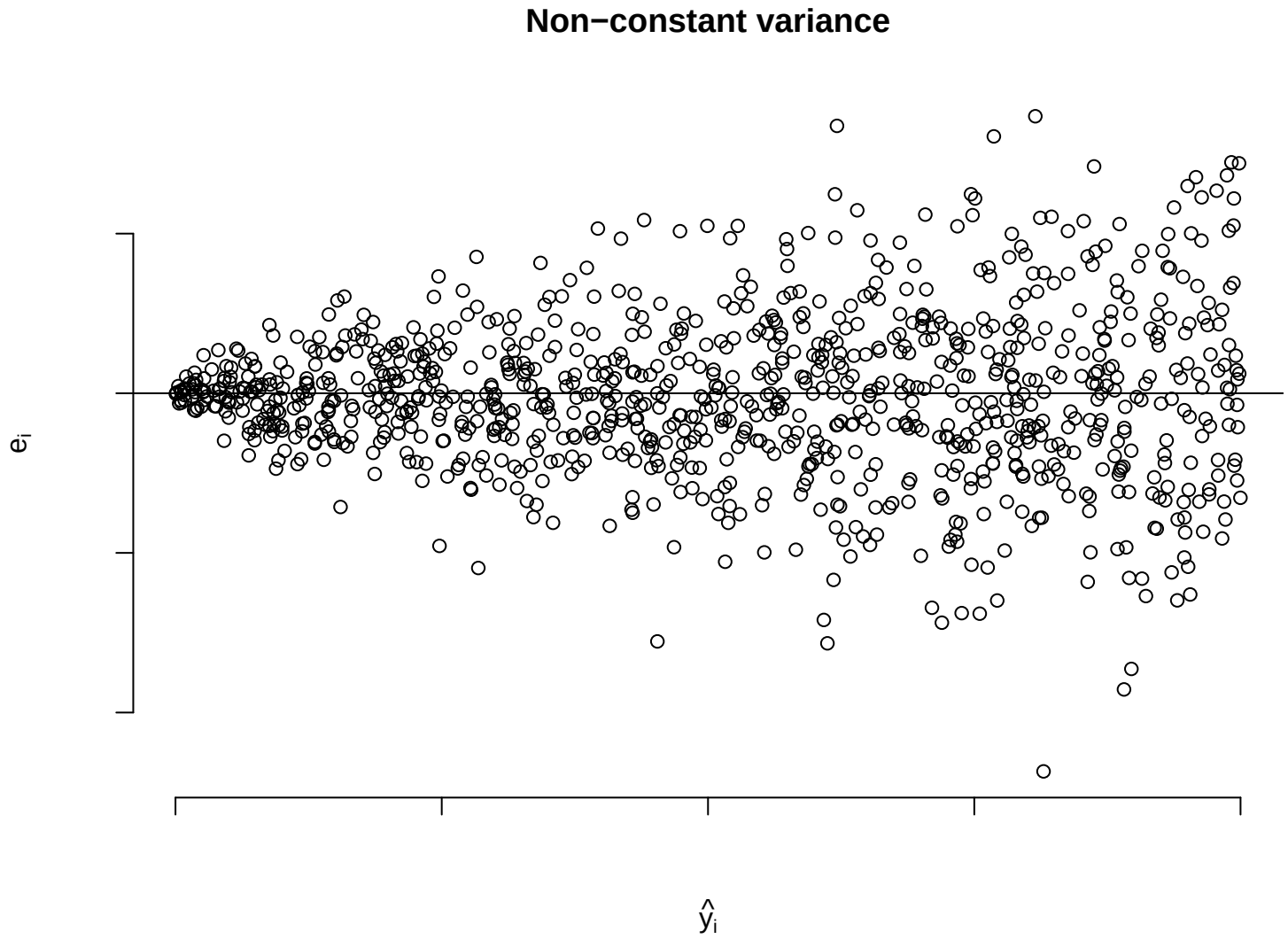
Scatterplots

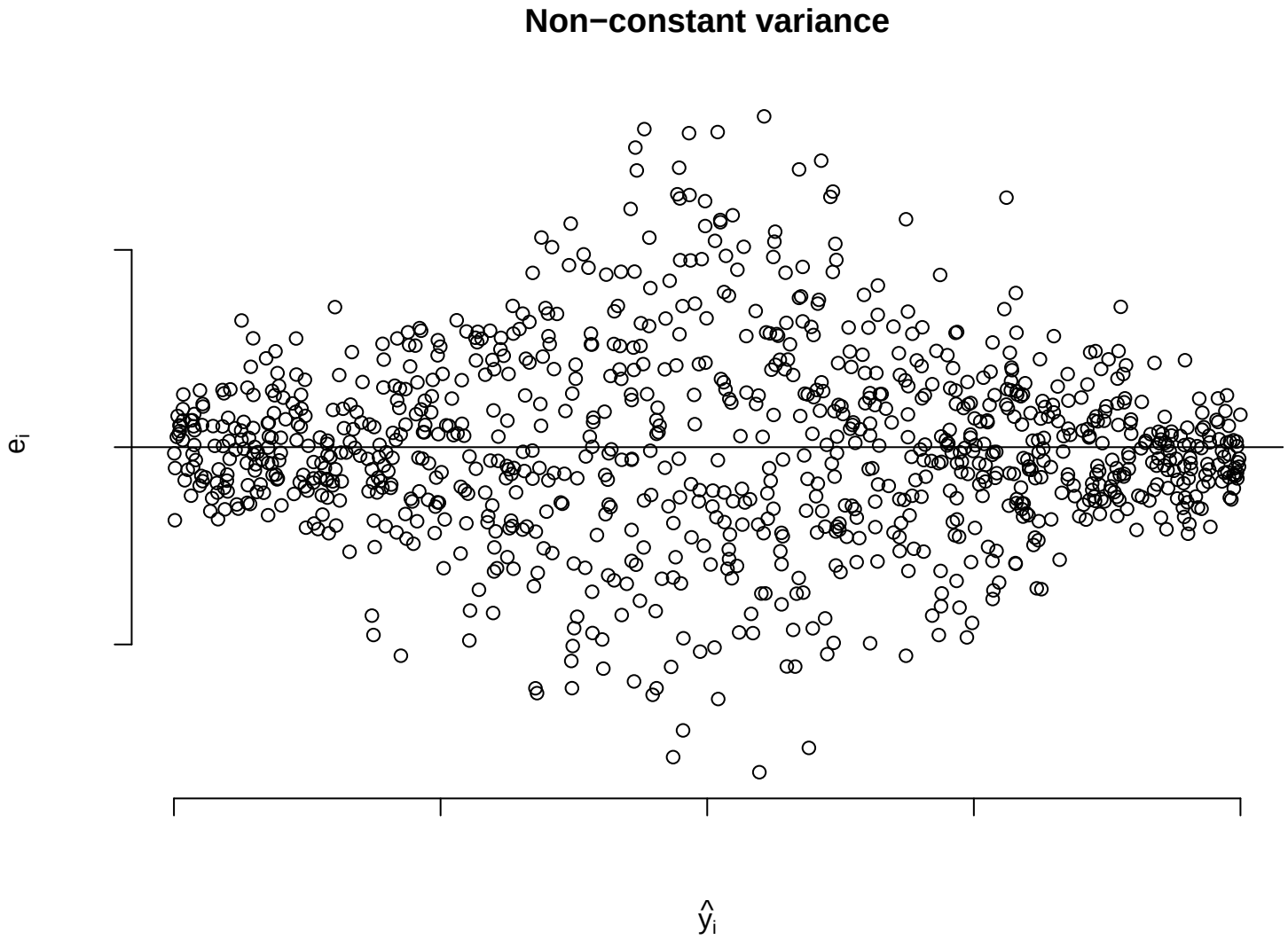
Another useful aid for inspection is a scatterplot of the residuals against the fitted values and/or the predictors. These plots can help us identify:

- Non-constant variance
- Violation of the assumption of linearity
- Potential outliers

Satisfactory residual plot




Non−constant variance
e_i
$\hat{y}_i$


Non−constant variance
e_i
$\hat{y}_i$

Example

Suppose the true relationship between a predictor, x, and an outcome, y is

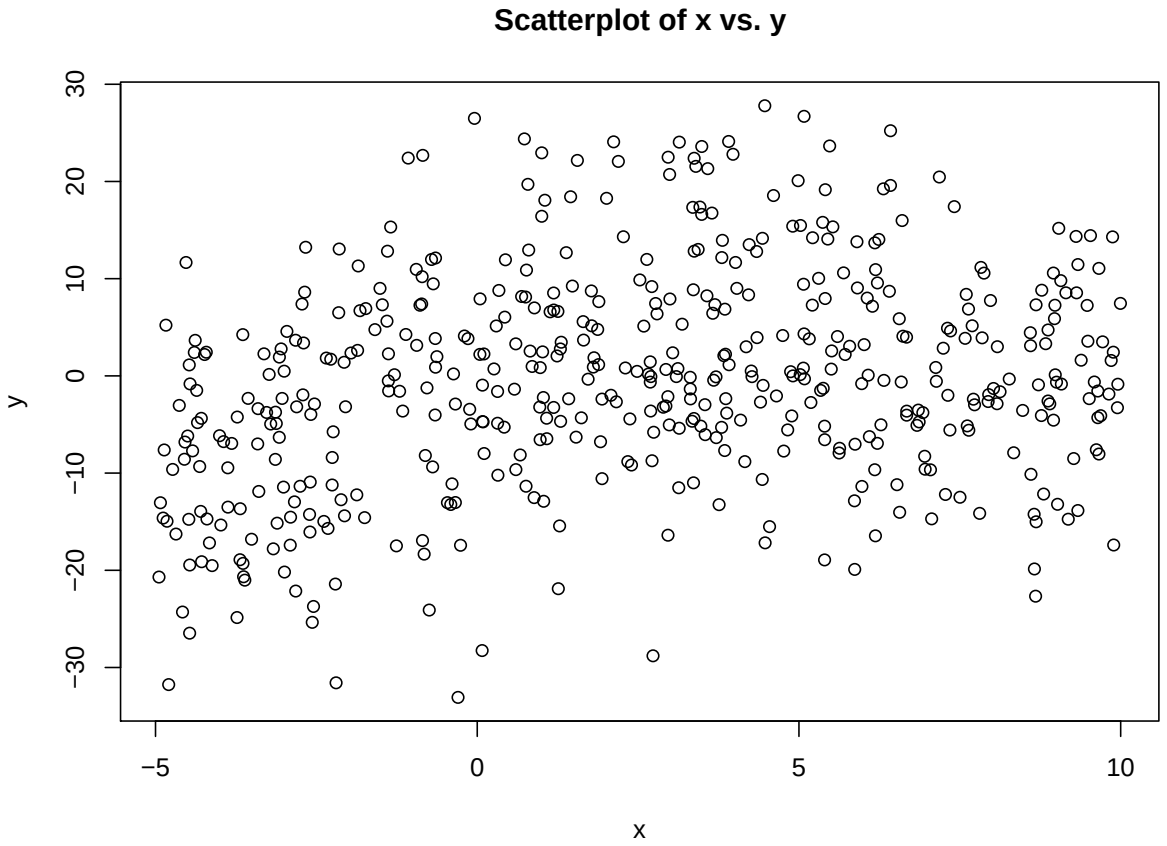
$$E(y_i) = 1 + 2x_i - 0.25x_i^2,$$

but we fit the model

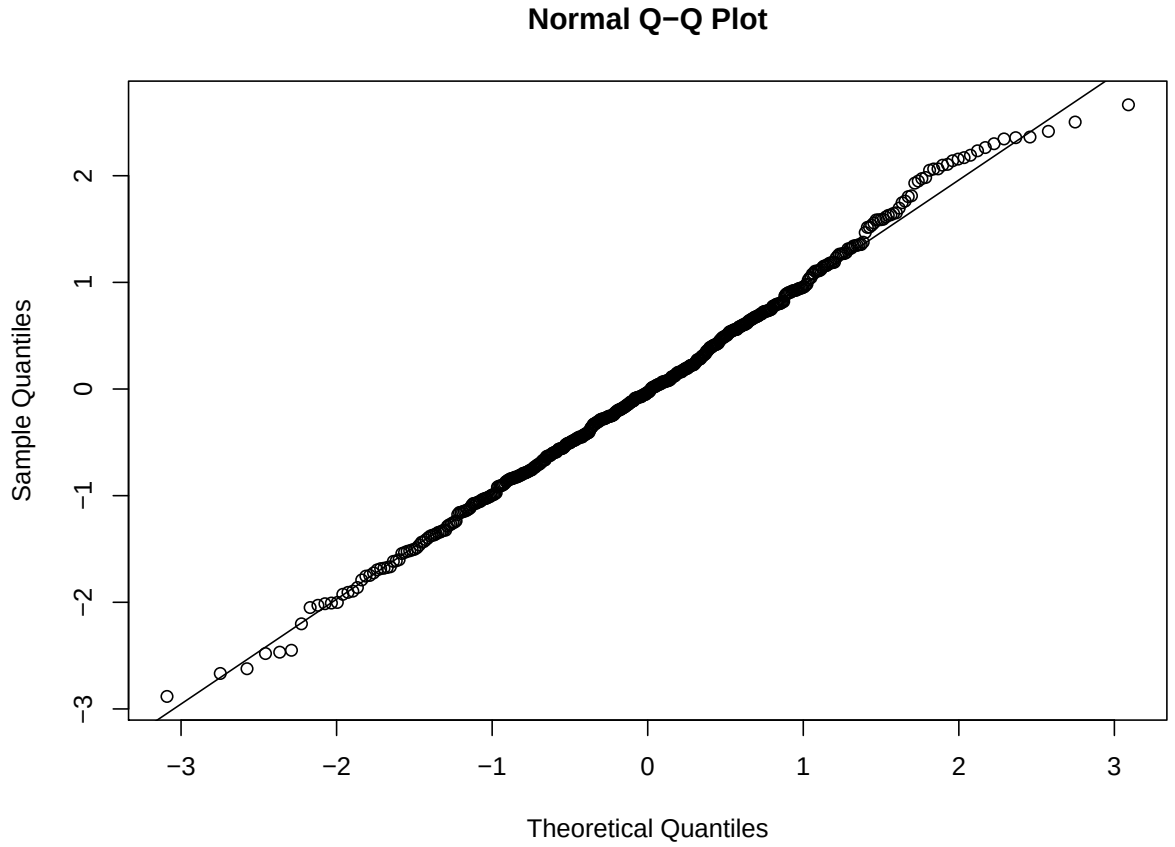
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Can we diagnose this with residual plots?

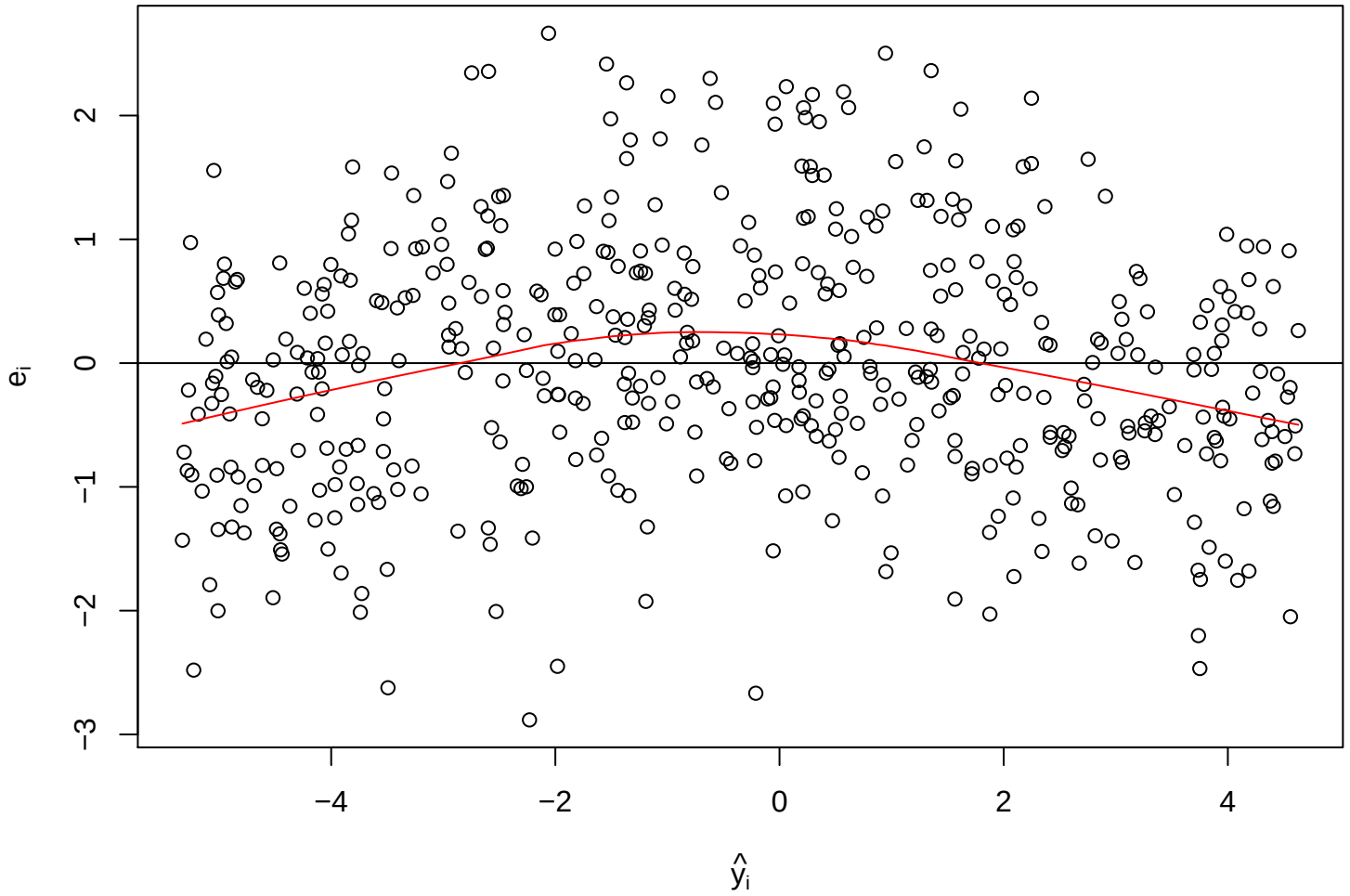
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−1.7761	0.5618	−3.16	0.0017
x	0.7569	0.1111	6.81	0.0000



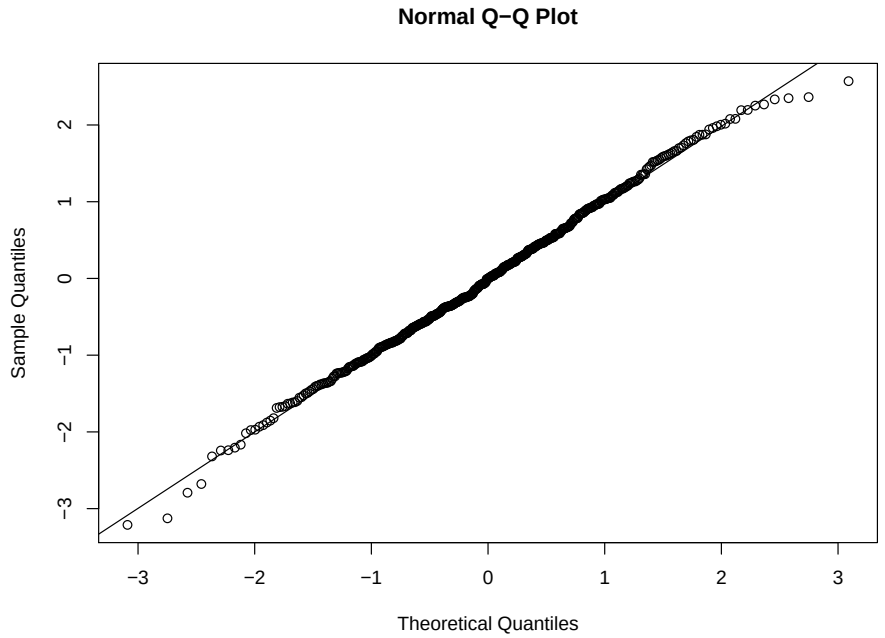
Fitted values versus residuals



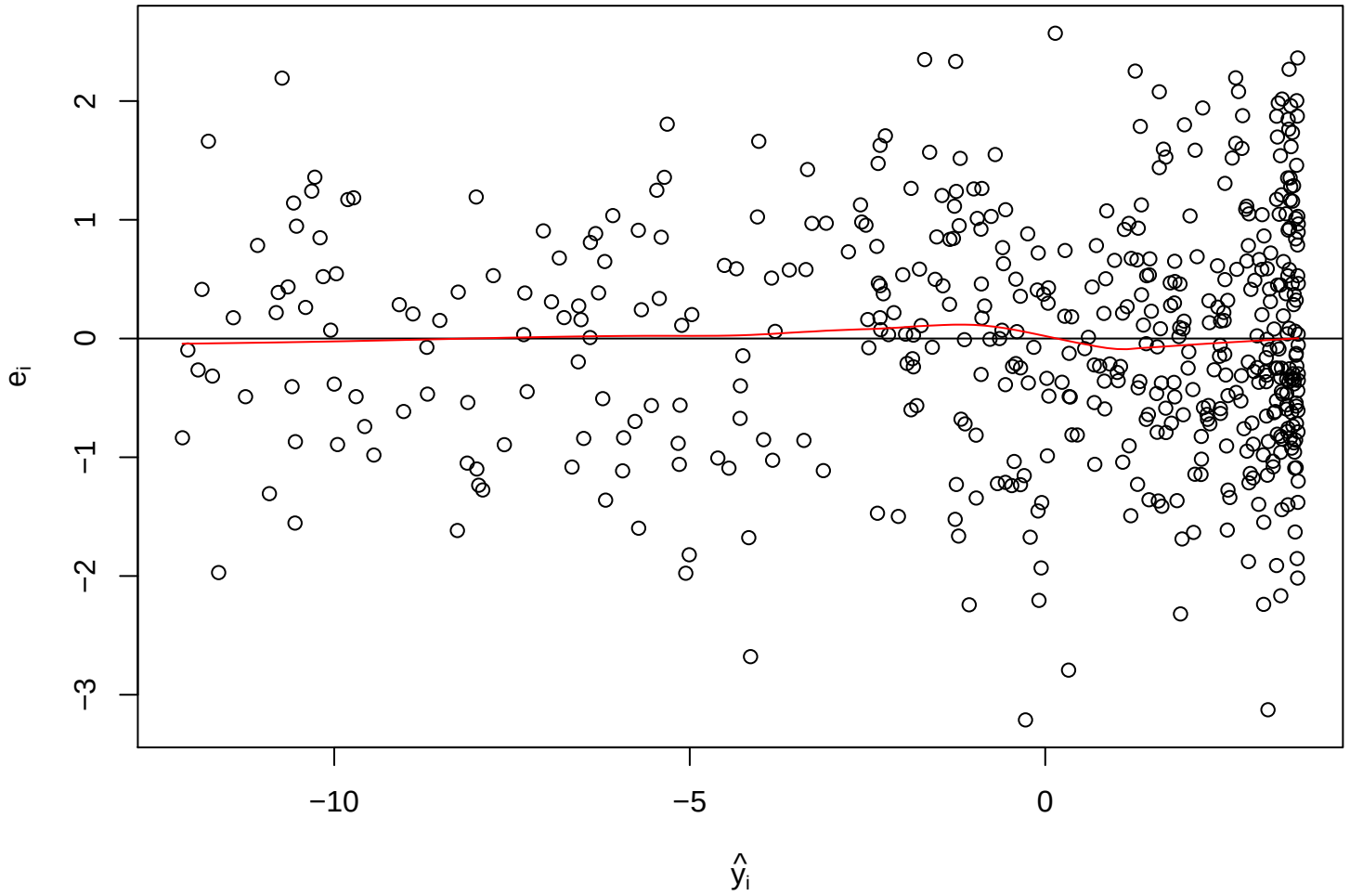
Next, we fit

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9824	0.6000	3.30	0.0010
x	1.9653	0.1637	12.00	0.0000
x2	−0.2640	0.0268	−9.85	0.0000



Fitted values versus residuals



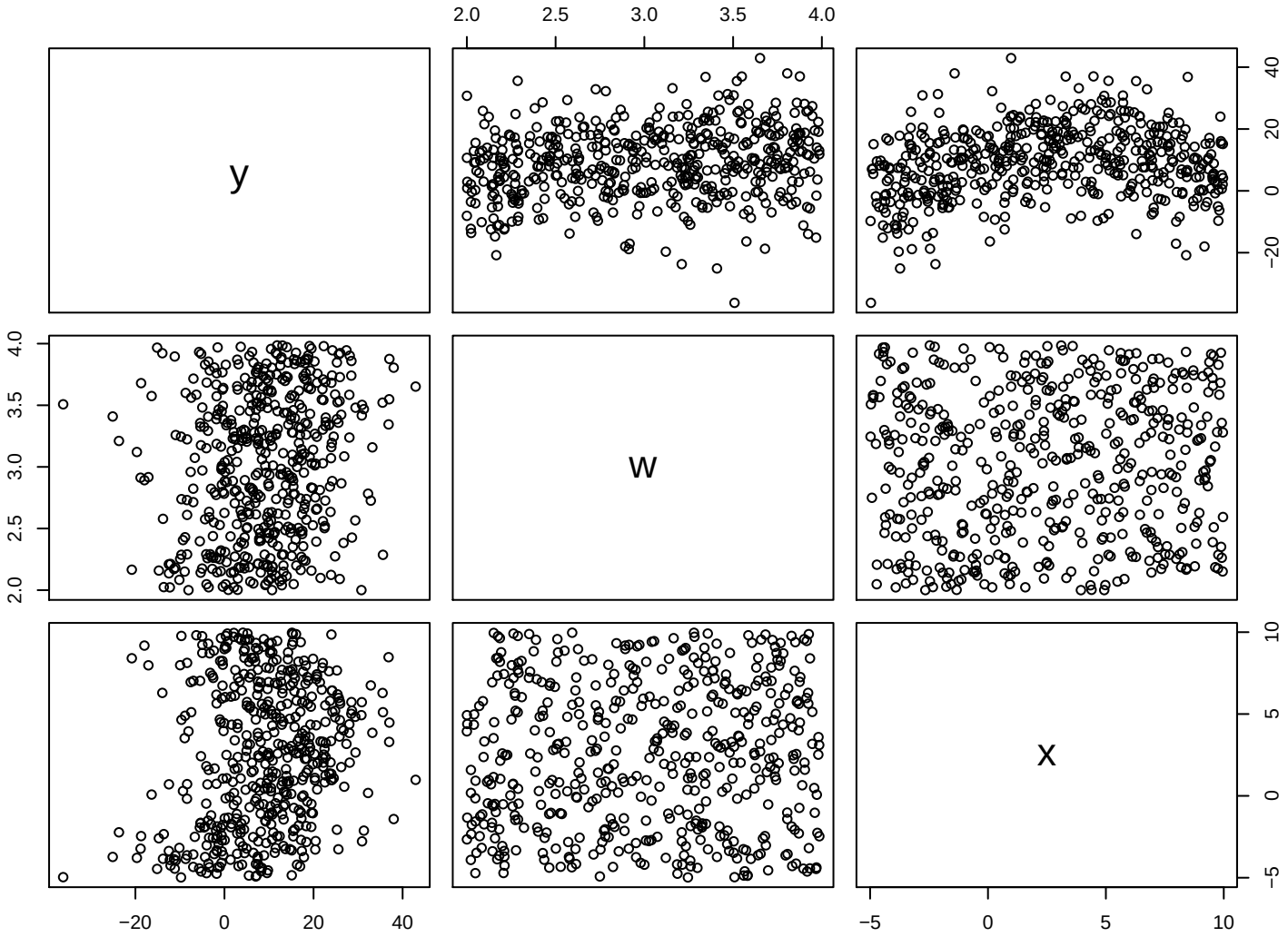
What if we have more than one predictor and only one is misspecified?
So, the true model is

$$E[y_i] = 1 + 3w_i + 2x_i - 0.25x_i^2$$

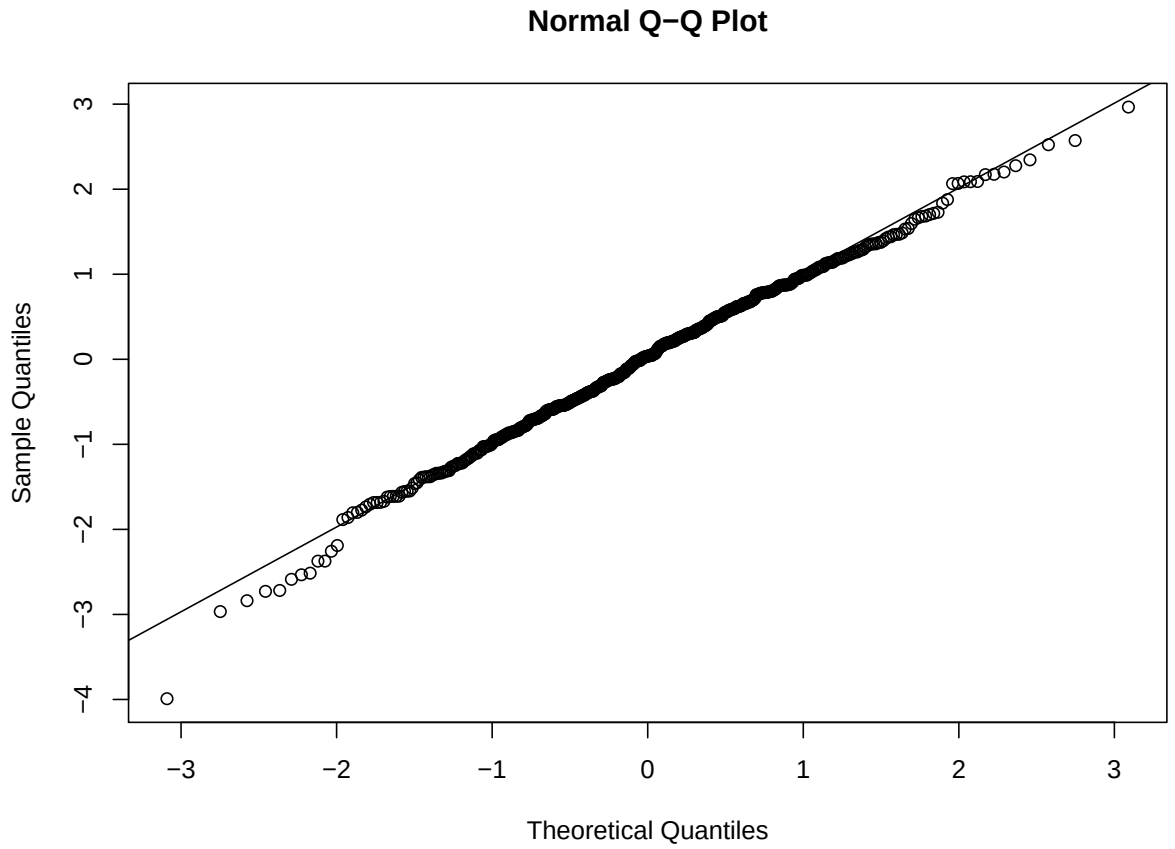
and we fit

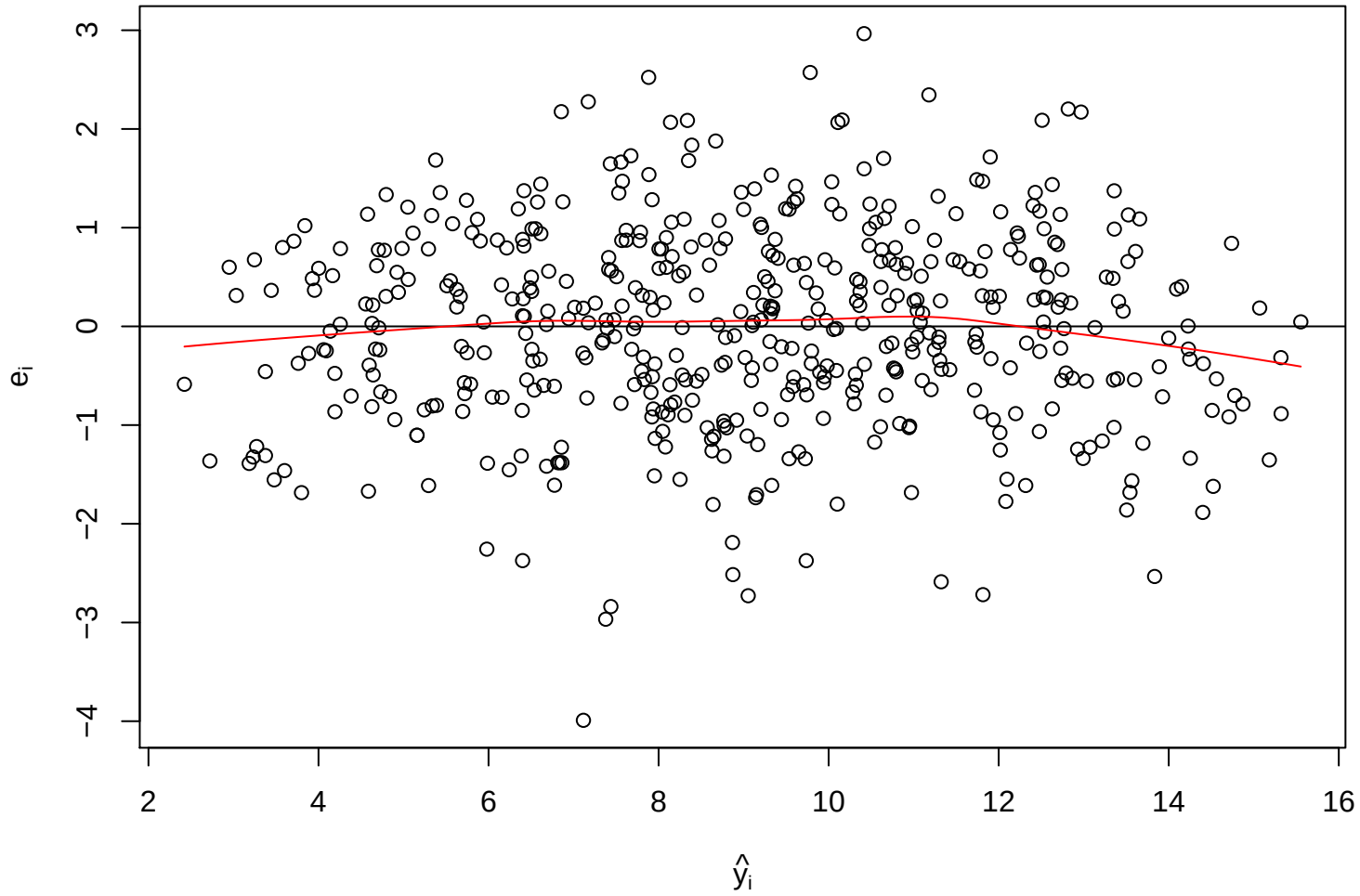
$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \epsilon_i.$$

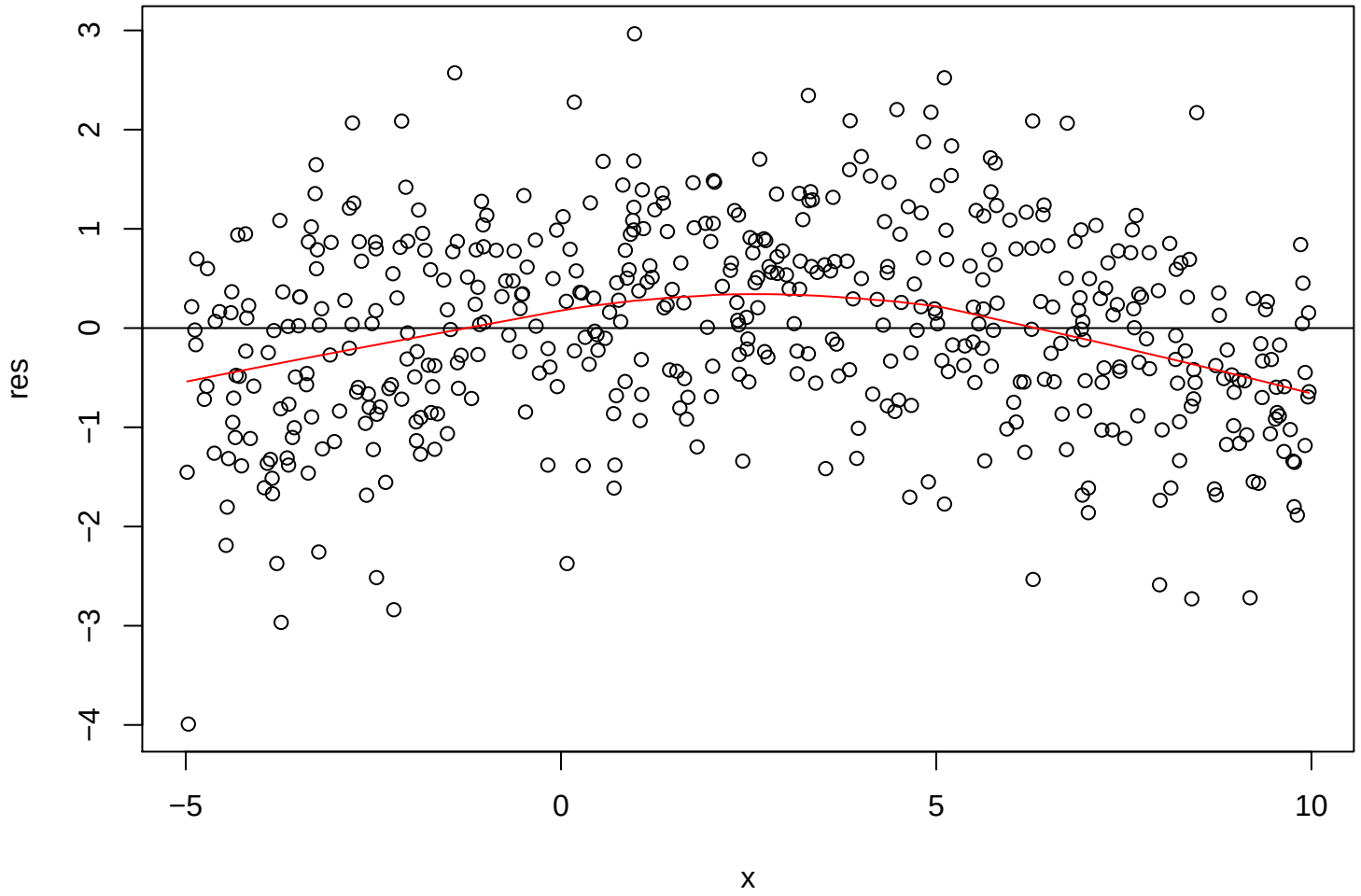
How can we diagnose model misspecification with residual plots in this case?

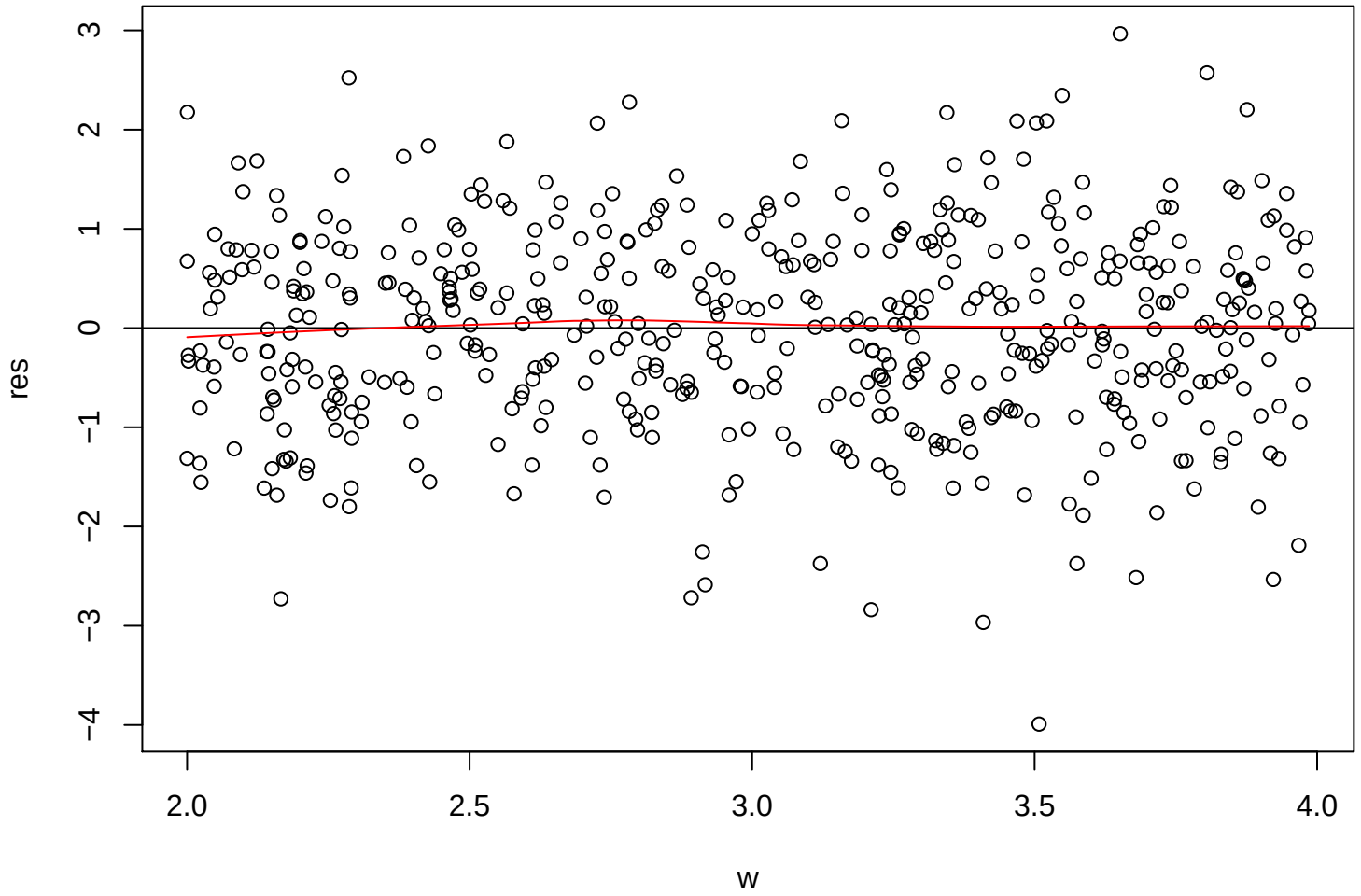

y
w
x
2.0
2.5
3.0
3.5
4.0
−20
0
20
40
−5
0
5
10

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−2.9863	2.6128	−1.14	0.2536
x	1.1326	0.1137	9.96	0.0000
w	2.9216	0.8532	3.42	0.0007



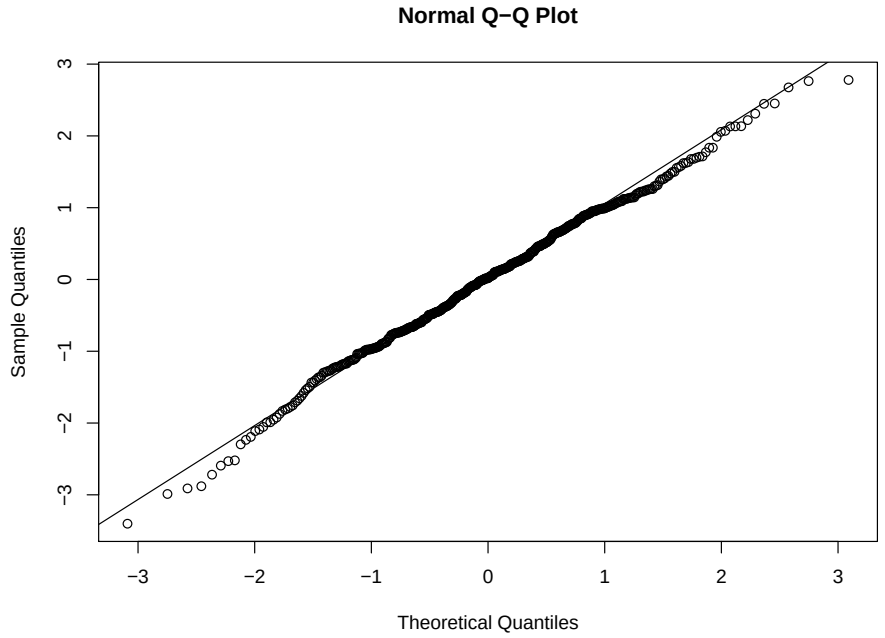

3
2
1
0
−1
−2
−3
−4
e_i
2
4
6
8
10
12
14
16
$\hat{y}_i$

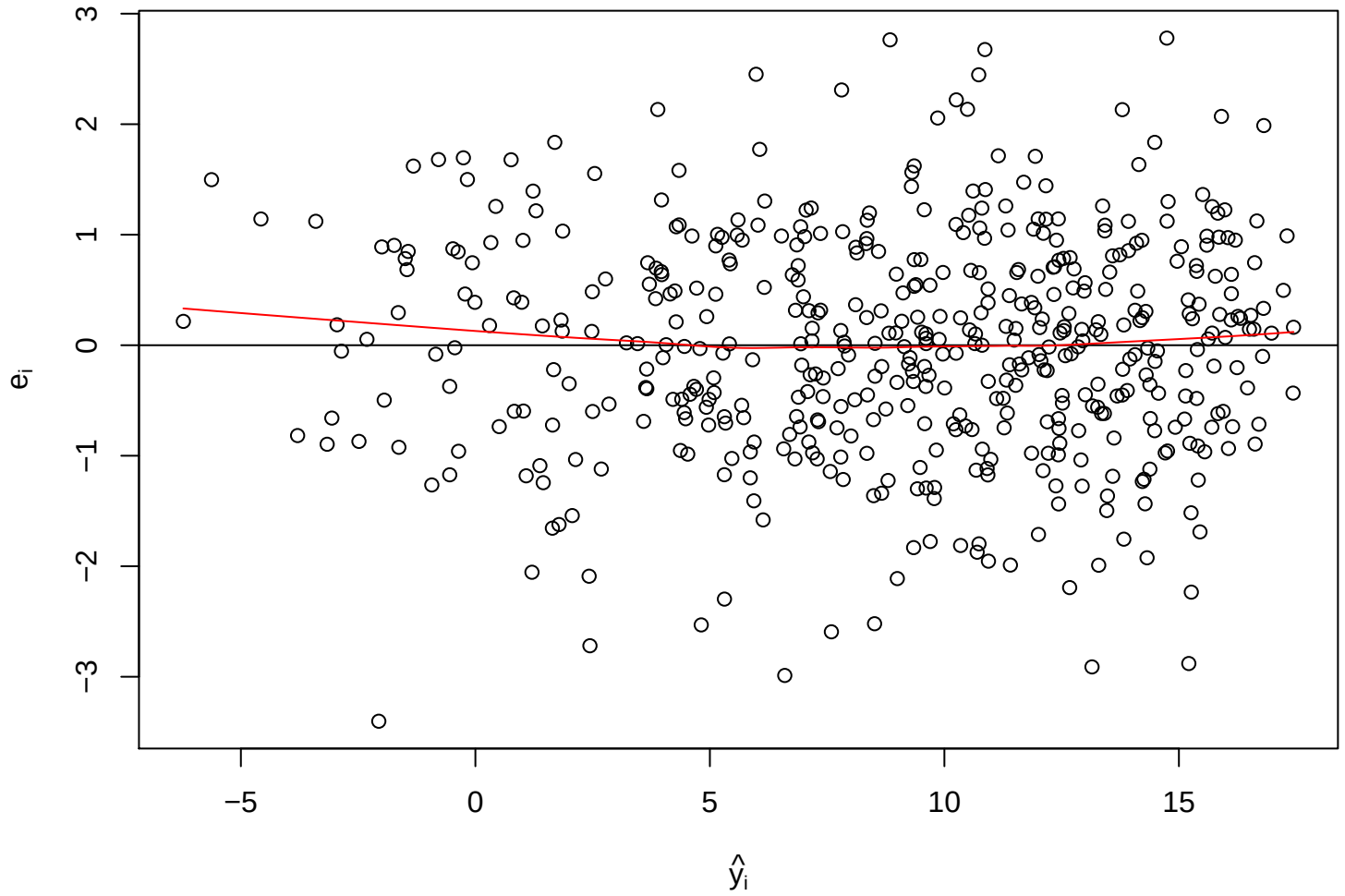

res
3
2
1
0
−1
−2
−3
−4
−5
0
5
10
x


res
3
2
1
0
−1
−2
−3
−4
2.0
2.5
3.0
3.5
4.0
w

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + \beta_3 x_i^2 \epsilon_i$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−1.1002	2.3876	−0.46	0.6452
x	2.2203	0.1603	13.85	0.0000
x2	−0.2884	0.0252	−11.43	0.0000
w	4.1124	0.7668	5.36	0.0000




3
2
1
0
−1
−2
−3
e_i
−5
0
5
10
15
$\hat{y}_i$

Partial regression plots

- A variation of the plot of residuals versus predictors
- Allows us to study the marginal relationship of a regressor given the other variables that are in the model.
- Also called the **added variable plot** or the **adjusted variable plot**

For example, suppose we are considering the model

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

and we are concerned about the correct specification of the relationship between x_1 and y_1. To create a partial regression plot, we

1. Regress y on x_2 and obtain the fitted values and residuals

$$\hat{y}_i(x_2) = \hat{\theta}_0 + \hat{\theta}_1 x_{i2}$$

$$e_i(y|x_2) = y_i - \hat{y}_i(x_2),\ i = 1, \ldots, n$$

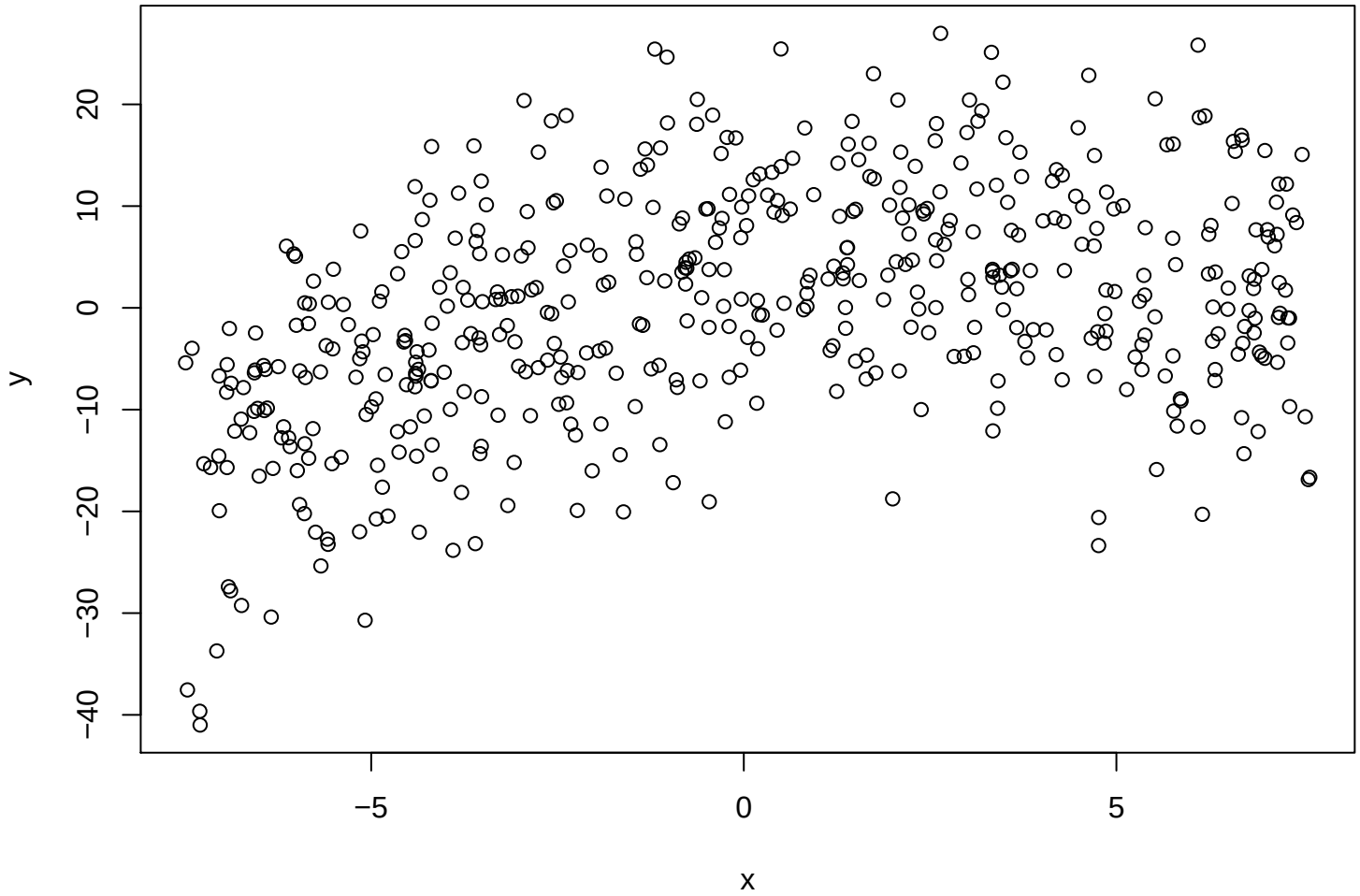
2. Regress x_1 on x_2 and calculate the residuals

$$\hat{x_{i1}}(x_2) = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2}$$

$$e_i(x_1|x_2) = x_{i1} - \hat{x_{i1}}(x_2),\ i = 1, \ldots, n$$

3. Plot the y residuals $e_i(y|x_2)$ against the x_1 residuals $e_i(x_1|x_2)$.

Partial regression plot from previous example



Comments on partial regression plots

- Use with caution - they only suggest possible relationships between the predictor and the response.
- In general, they will not detect interactions between regressors.
- The presence of strong multicollinearity can cause partial regression plots to give incorrect information.

Next

- Variance stabilizing transformations
- Identifying outliers and influential observations.