Applied Regression Modeling:
A Business Approach
Chapter 3: Multiple Linear Regression
Sections 3.4–3.6

by Iain Pardoe

Regression model assumptions

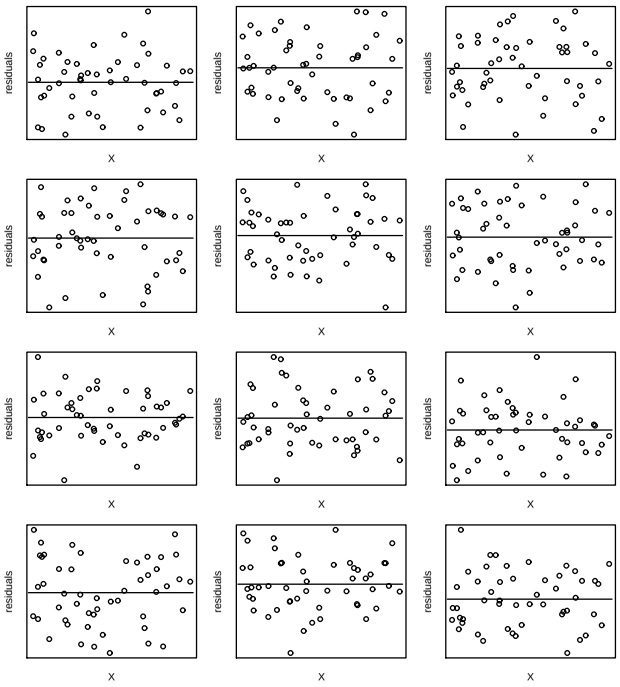
Four assumptions about random errors, $e = Y - \mathrm{E}(Y) = Y - b_0 - b_1 X_1 - \cdots - b_k X_k$:

- Probability distribution of e at each set of values $(X_1, X_2, \ldots, X_k)$ has a **mean of zero**;
- Probability distribution of e at each set of values $(X_1, X_2, \ldots, X_k)$ has **constant variance**;
- Probability distribution of e at each set of values $(X_1, X_2, \ldots, X_k)$ is **normal**;
- Value of e for one observation is **independent** of the value of e for any other observation.

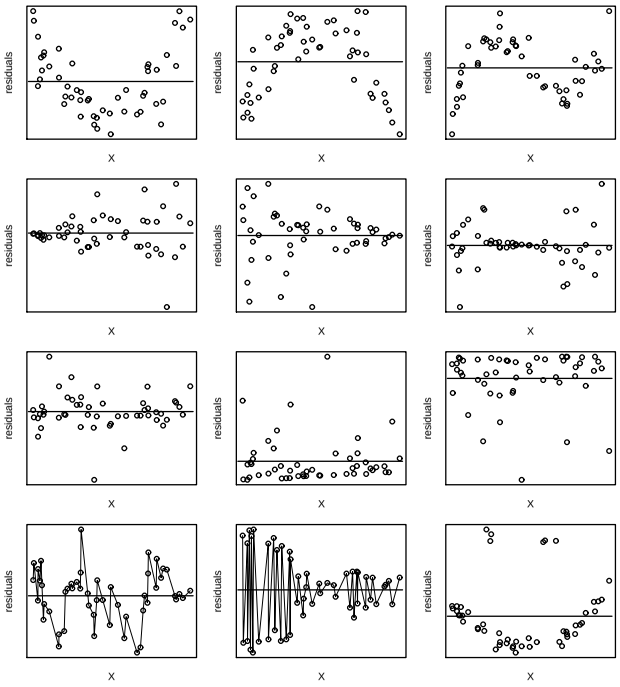
Checking the model assumptions

- Calculate residuals, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X_1 - \cdots - \hat{b}_k X_k$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and a function of $(X_1, X_2, \ldots, X_k)$ along the horizontal axis (e.g., $\hat{Y}$ or one of the X's).
  - Assess **zero mean** assumption—do the residuals average out to zero as we move across the plot from left to right?
  - Assess **constant variance** assumption—is the (vertical) variation of the residuals similar as we move across the plot from left to right?
  - Assess **independence** assumption—do residuals look "random" with no systematic patterns?
- Draw a histogram and QQ-plot of the residuals.
  - Assess **normality** assumption—does histogram look approximately bell-shaped and symmetric and do QQ-plot points lie close to line?

Residual plots which pass

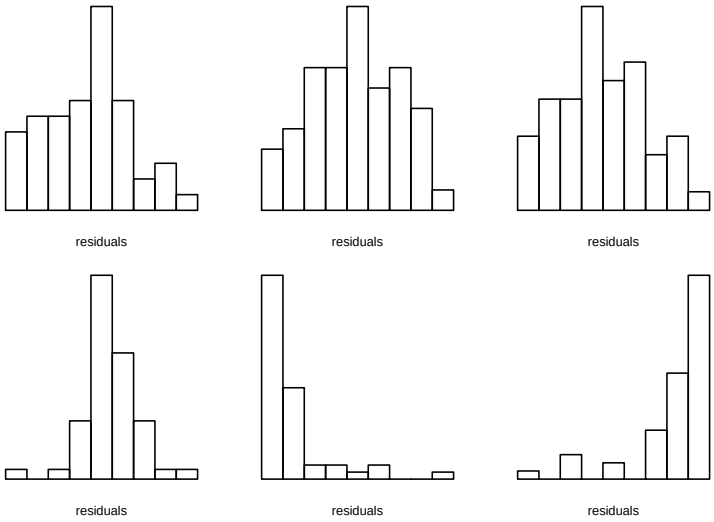


Residual plots which fail



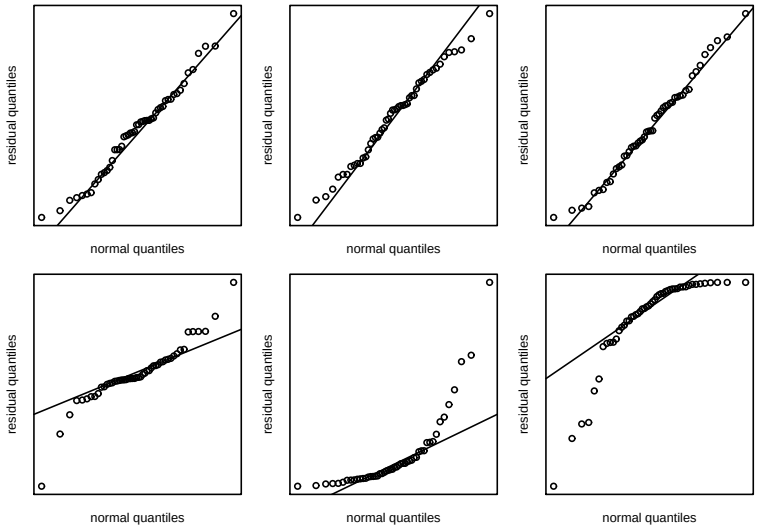
Histograms of residuals

Upper three pass, lower three fail



QQ-plots of residuals

Upper three pass, lower three fail

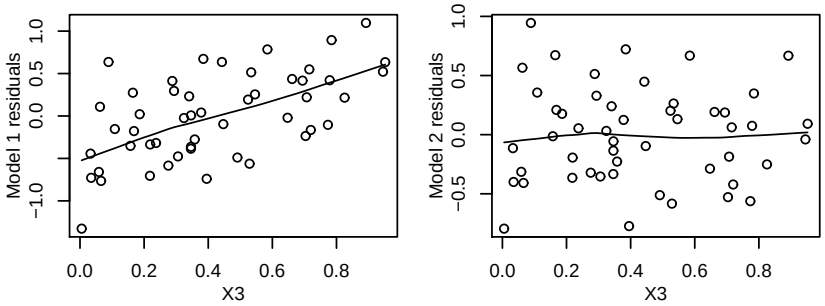


Assessing assumptions in practice

- Assessing assumptions in practice can be difficult and time-consuming.
- Taking the time to check the assumptions is worthwhile and can provide additional support for any modeling conclusions.
- *Clear* violation of one or more assumptions could mean results are questionable and should probably not be used.
- Possible remedy: try a different subset of available predictors (further ideas to come in Chapter 4).
- Regression results tend to be quite robust to *mild* violations of assumptions.
- Checking assumptions when n is very small (or very large) can be particularly challenging.
- Example: **MLRA** data file.

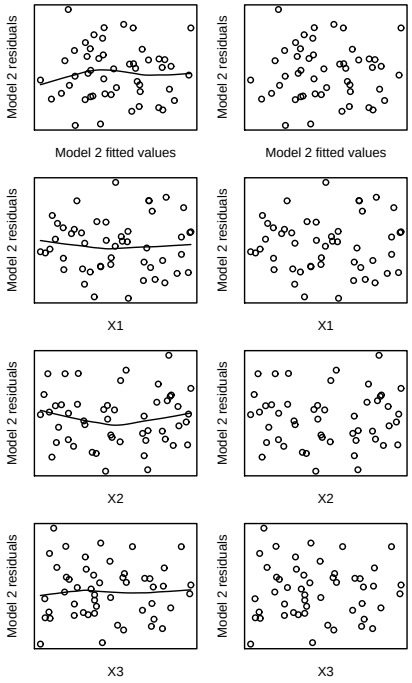
MLRA residual plots—zero mean check

Model 1 on the left: $\mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2$.
Model 2 on the right: $\mathrm{E}(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3$.



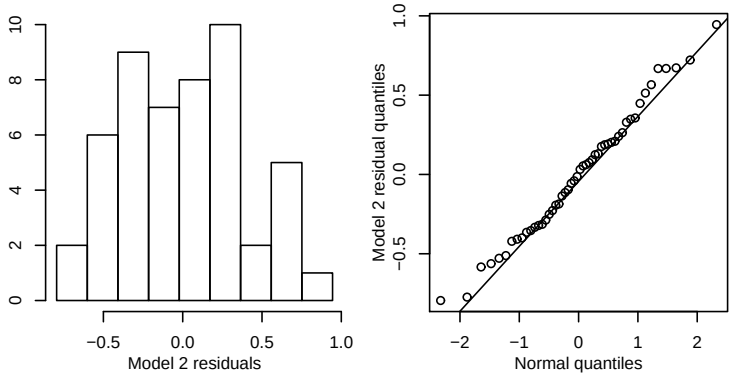
Plots include "loess fitted lines" (computational method for applying "slicing/averaging" technique).
Do either of the models fail the zero mean assumption?

MLRA model 2 residual plots



MLRA residual histogram and QQ-plot

The approximately bell-shaped and symmetric histogram and QQ-plot points lying close to the line support the normality assumption.



Shipping example model building

Model Summary

Model	R Squared	Adjusted R Squared	Regression Std. Error	Change Statistics F-stat	df1	df2	Pr(>F)
1	0.808 [a]	0.786	8.815				
2	0.820 [b]	0.771	9.103	0.472	2	15	0.633

[a] Predictors: (Intercept), X1, X3.

[b] Predictors: (Intercept), X1, X2, X3, X4.

There is no evidence at the 5% significance level that X_2 (proportion shipped by truck) or X_4 (week) provide useful information about Y (weekly labor hours) beyond the information provided by X_1 (total weight shipped in thousands of pounds) and X_3 (average shipment weight in pounds).

Shipping example two-predictor model results

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.899 [a]	0.808	0.786	8.815

[a] Predictors: (Intercept), X1, X3.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
1	(Intercept)	110.431	24.856	4.443	0.000
	X1	5.001	2.261	2.212	0.041
	X3	−2.012	0.668	−3.014	0.008

95% Confidence Interval

Model	Lower Bound	Upper Bound
X1	0.231	9.770
X3	−3.420	−0.604

[a] Response variable: Y.

Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between Y and X_1 (holding X_3 constant)
  and between Y and X_3 (holding X_1 constant).
- Estimated equation: $\hat{Y} = 110.43 + 5.00X_1 - 2.01X_3$.
- $X_1 = X_3 = 0$ makes no sense for this application, nor do we have data close to $X_1 = X_3 = 0$, so cannot meaningfully interpret $\hat{b}_0 = 110.43$.
- Expect increase of 5 weekly labor hours when total weight increases 1000 pounds and ave. shipment weight remains constant, for total weights of 2000–$10{,}000$ pounds and ave. weights of 10–30 pounds (95% confident increase is 0.23–9.77).
- Expect decrease of 2.01 weekly labor hours when ave. weight increases 1 pound and total weight remains constant, for total weights of 2000–$10{,}000$ pounds and ave. weights of 10–30 pounds (95% confident decrease is 0.60–3.42).

Interpreting model results (cont)

- Can expect a prediction of unobserved weekly labor hours from particular values of total weight shipped and average shipment weight to be accurate to within approximately ± 17.6 (with 95% confidence).
- 80.8% of the variation in weekly labor hours (about its mean) can be explained by a multiple linear regression relationship between labor hours and (total weight shipped, average shipment weight).

Confidence interval for population mean, E(Y)

- Estimate the mean (or expected) value of Y at particular values of $(X_1, X_2, \ldots, X_k)$.
- Formula: $\hat{Y} \pm$ t-percentile$(s_{\hat{Y}})$.
- Interval is narrower:
  - when n is large;
  - when X's are close to their sample means;
  - when the regression standard error, s, is small;
  - for lower levels of confidence.
- Example: for shipping example two-predictor model, the 95% confidence interval for $\mathrm{E}(Y)$ when $X_1 = 6$ and $X_3 = 20$ is $(95.4,\ 105.0)$.
- Interpretation: we're 95% confident that expected weekly labor hours is between 95.4 and 105.0 when total weight shipped is 6000 pounds and average shipment weight is 20 pounds.

Prediction interval for an individual Y-value

- Predict an individual value of Y at particular values of $(X_1, X_2, \ldots, X_k)$.
- Formula: $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$.
- Interval is narrower:
  - when n is large;
  - when X's are close to their sample means;
  - when the regression standard error, s, is small;
  - for lower levels of confidence.
- Since $s_{\hat{Y}^*} > s_{\hat{Y}}$, prediction interval is wider than confidence interval.
- Example: for shipping example two-predictor model, the 95% prediction interval for Y^* when $X_1 = 6$ and $X_3 = 20$ is $(81.0,\ 119.4)$.
- Interpretation: we're 95% confident that actual labor hours in a week is between 81.0 and 119.4 when total weight shipped is 6000 pounds and average shipment weight is 20 pounds.

