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

by Iain Pardoe
Regression model assumptions

3.4 Model assumptions

Four assumptions about random errors,
\[ e = Y - E(Y) = Y - b_0 - b_1 X_1 - \cdots - b_k X_k. \]
Four assumptions about random errors,
\[ e = Y - \mathbb{E}(Y) = Y - b_0 - b_1 X_1 - \cdots - b_k X_k: \]

- Probability distribution of \( e \) at each set of values \((X_1, X_2, \ldots, X_k)\) has a **mean of zero**;
Four assumptions about random errors,
\[ e = Y - \mathbb{E}(Y) = Y - b_0 - b_1 X_1 - \cdots - b_k X_k: \]

- 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**;
Four assumptions about random errors, 
\[ e = Y - E(Y) = Y - b_0 - b_1X_1 - \cdots - b_kX_k; \]

- 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**;
Regression model assumptions

Four assumptions about random errors,
\[ e = Y - E(Y) = Y - b_0 - b_1 X_1 - \cdots - b_k X_k : \]

- 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.
• Calculate residuals,
\[ \hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1X_1 - \cdots - \hat{b}_kX_k. \]

• 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).
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?
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?
3.4 Model assumptions

Regression model assumptions
Checking the model assumptions

Residual plots which pass
Residual plots which fail
Histograms of residuals
QQ-plots of residuals
Assessing assumptions in practice
MLRA residual plots—zero mean check
MLRA model 2 residual plots
MLRA residual histogram and QQ-plot

3.5 Model interpretation

3.6 Estimation and prediction

© Iain Pardoe, 2006

\[
\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X_1 - \cdots - \hat{b}_k X_k.
\]

- Calculate residuals,
- 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?
Checking the model assumptions

- Calculate residuals,
  \[ \hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1X_1 - \cdots - \hat{b}_kX_k. \]

- 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.
Checking the model assumptions

3.4 Model assumptions

Regression model assumptions

Checking the model assumptions

Residual plots which pass
Residual plots which fail
Histograms of residuals
QQ-plots of residuals
Assessing assumptions in practice
MLRA residual plots—zero mean check
MLRA model 2 residual plots
MLRA residual histogram and QQ-plot

3.5 Model interpretation

3.6 Estimation and prediction

© Iain Pardoe, 2006

- 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?
3.4 Model assumptions
Regression model assumptions
Checking the model assumptions

Residual plots which pass
Residual plots which fail
Histograms of residuals
QQ-plots of residuals
Assessing assumptions in practice
MLRA residual plots—zero mean check
MLRA model 2 residual plots
MLRA residual histogram and QQ-plot

3.5 Model interpretation

3.6 Estimation and prediction

© Iain Pardoe, 2006
3.4 Model assumptions
Regression model assumptions
Checking the model assumptions
Residual plots which pass

Residual plots which fail
Histograms of residuals
QQ-plots of residuals
Assessing assumptions in practice
MLRA residual plots—zero mean check
MLRA model 2 residual plots
MLRA residual histogram and QQ-plot

3.5 Model interpretation

3.6 Estimation and prediction

© Iain Pardoe, 2006
Upper three pass, lower three fail

Histograms of residuals

QQ-plots of residuals

Assessing assumptions in practice

MLRA residual plots—zero mean check

MLRA model 2 residual plots

MLRA residual histogram and QQ-plot
3.4 Model assumptions
Regression model assumptions
Checking the model assumptions
Residual plots which pass
Residual plots which fail
Histograms of residuals
QQ-plots of residuals
Assessing assumptions in practice
MLRA residual plots—zero mean check
MLRA model 2 residual plots
MLRA residual histogram and QQ-plot

3.5 Model interpretation

3.6 Estimation and prediction

© Iain Pardoe, 2006
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.
Model 1 on the left: $E(Y) = b_0 + b_1 X_1 + b_2 X_2$.
Model 2 on the right: $E(Y) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$.

Plots include “loess fitted lines” (computational method for applying “slicing/averaging” technique). Do either of the models fail the zero mean assumption?
3.4 Model assumptions
Regression model assumptions
Checking the model assumptions
Residual plots which pass
Residual plots which fail
Histograms of residuals
QQ-plots of residuals
Assessing assumptions in practice
MLRA residual plots—zero mean check

3.5 Model interpretation

3.6 Estimation and prediction

MLRA model 2 residual plots
MLRA residual histogram and QQ-plot

© Iain Pardoe, 2006
The approximately bell-shaped and symmetric histogram and QQ-plot points lying close to the line support the normality assumption.
Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R Squared</th>
<th>Adjusted R Squared</th>
<th>Std. Error</th>
<th>F-stat</th>
<th>df1</th>
<th>df2</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.808</td>
<td>0.786</td>
<td>8.815</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.820</td>
<td>0.771</td>
<td>9.103</td>
<td>0.472</td>
<td>2</td>
<td>15</td>
<td>0.633</td>
</tr>
</tbody>
</table>

Predictors: (Intercept), X1, X3.

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Multiple R</th>
<th>R Squared</th>
<th>Adjusted R Squared</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.899&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.808</td>
<td>0.786</td>
<td>8.815</td>
</tr>
</tbody>
</table>

<sup>a</sup> Predictors: (Intercept), X1, X3.

### Parameters<sup>a</sup>

| 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>0.231</td>
<td>9.770</td>
</tr>
<tr>
<td>X3</td>
<td>-3.420</td>
<td>-0.604</td>
</tr>
</tbody>
</table>

<sup>a</sup> Response variable: Y.
• We found a statistically significant straight-line relationship (at a 5% significance level) between $Y$ and $X_1$ (holding $X_3$ constant)
• 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).
• 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: $\hat{Y} = 110.43 + 5.00X_1 - 2.01X_3$. 
• 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 $\hat{b}_0 = 110.43$. 
• 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).
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).
• 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).
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).
• Estimate the mean (or expected) value of $Y$ at particular values of $(X_1, X_2, \ldots, X_k)$.
• Formula: $\hat{Y} \pm t\text{-percentile}(s_{\hat{Y}})$.
• 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.
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 $E(Y)$ 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: \( \hat{Y}^* \pm t\text{-percentile}(s_{\hat{Y}^*}) \).
- 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.
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.
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.