Applied Regression Modeling: A Business Approach
Chapter 2: Simple Linear Regression
Sections 2.4–2.7

by Iain Pardoe

Regression model assumptions

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

Regression model assumptions

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

- Probability distribution of e at each value of X has a **mean of zero**;

Regression model assumptions

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

- Probability distribution of e at each value of X has a **mean of zero**;
- Probability distribution of e at each value of X has **constant variance**;

Regression model assumptions

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

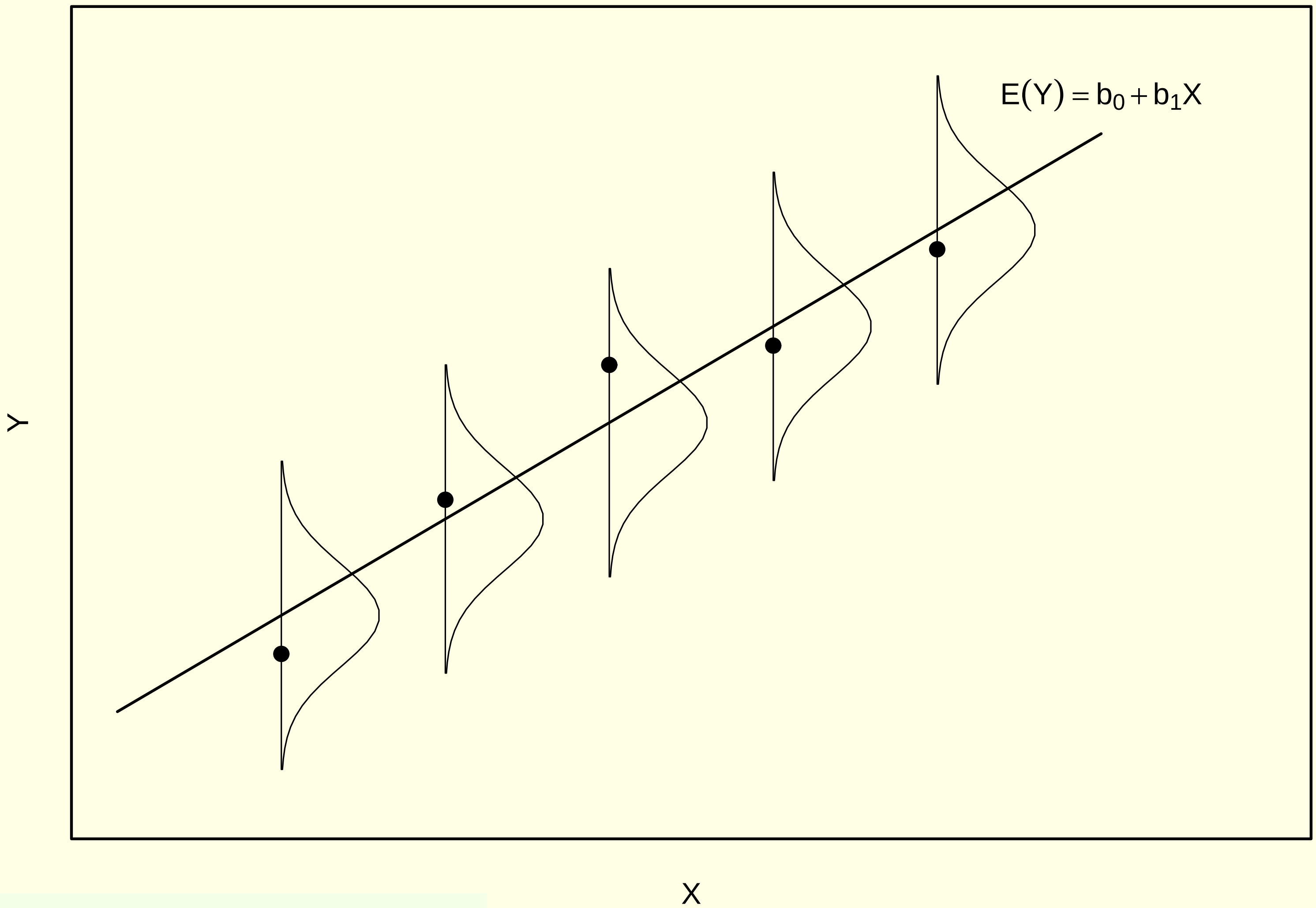
- Probability distribution of e at each value of X has a **mean of zero**;
- Probability distribution of e at each value of X has **constant variance**;
- Probability distribution of e at each value of X is **normal**;

Regression model assumptions

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

- Probability distribution of e at each value of X has a **mean of zero**;
- Probability distribution of e at each value of X has **constant variance**;
- Probability distribution of e at each value of X is **normal**;
- Value of e for one observation is **independent** of the value of e for any other observation.

Random error probability distributions.



Checking the model assumptions

- Calculate residuals, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and X along the horizontal axis.

Checking the model assumptions

- Calculate residuals, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and X along the horizontal axis.
    - 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, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and X along the horizontal axis.
  - 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?

Checking the model assumptions

- Calculate residuals, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and X along the horizontal axis.
    - 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?

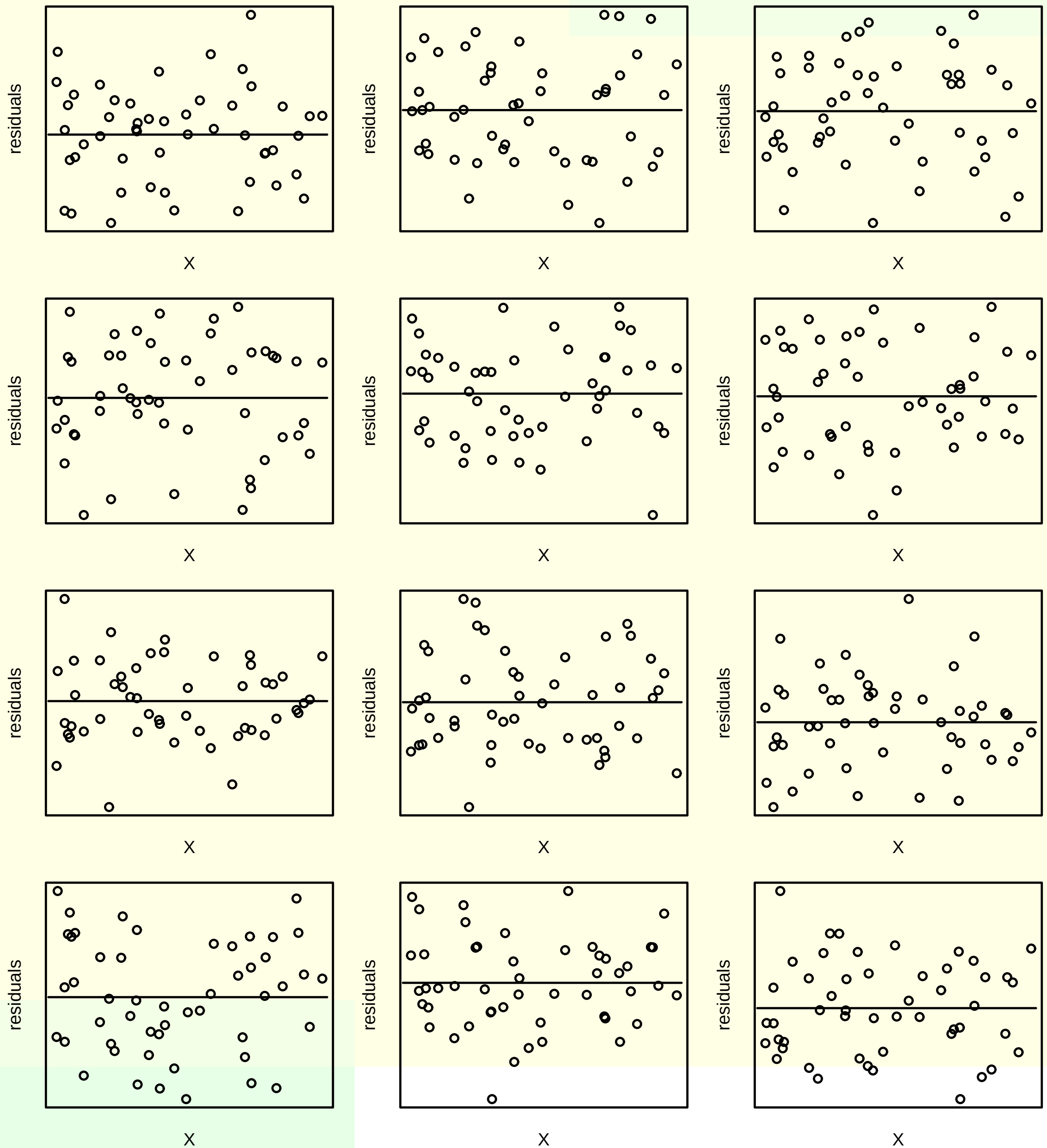
Checking the model assumptions

- Calculate residuals, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and X along the horizontal axis.
  - 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.

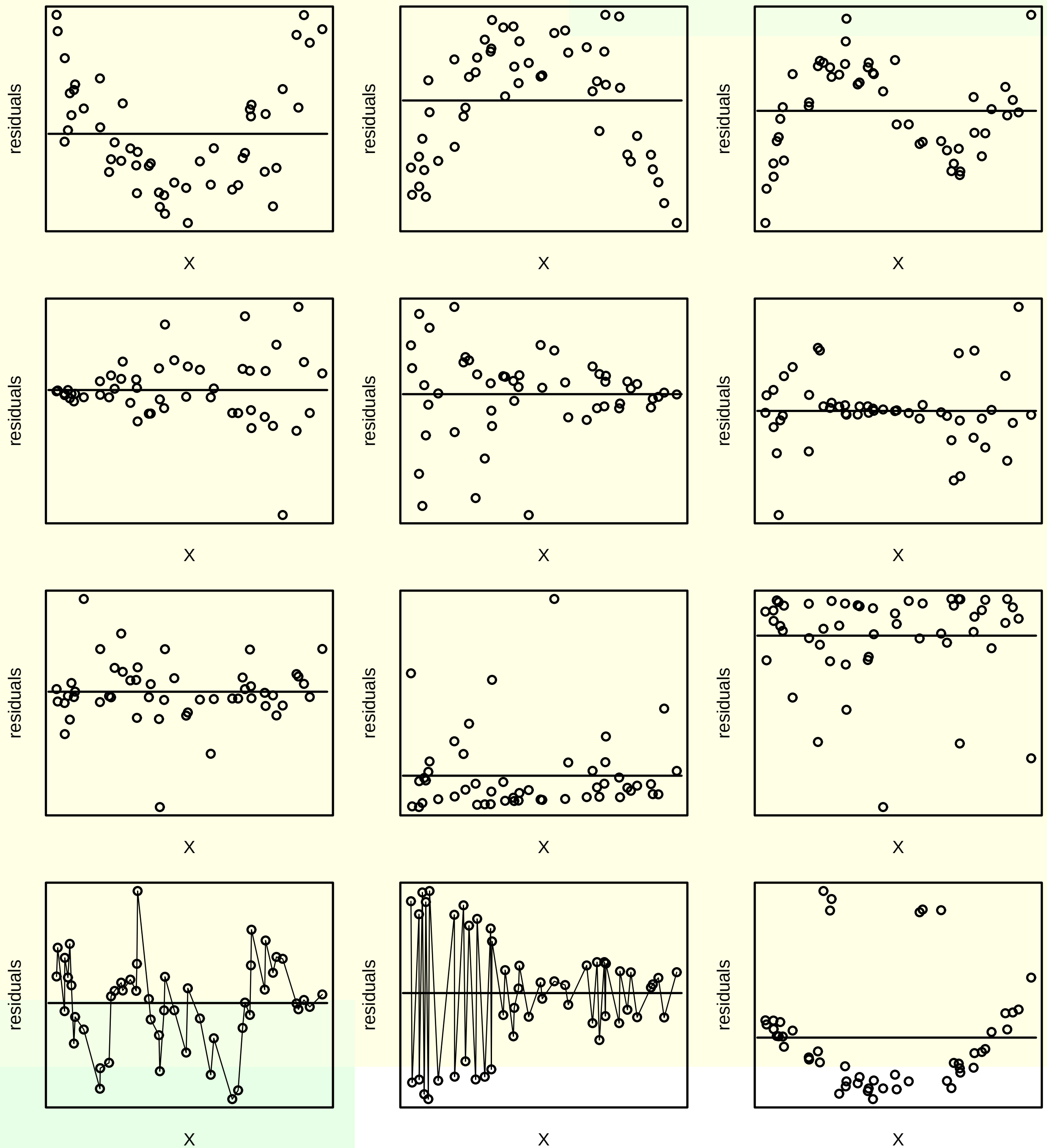
Checking the model assumptions

- Calculate residuals, $\hat{e} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$.
- Draw a residual plot with $\hat{e}$ along the vertical axis and X along the horizontal axis.
  - 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?

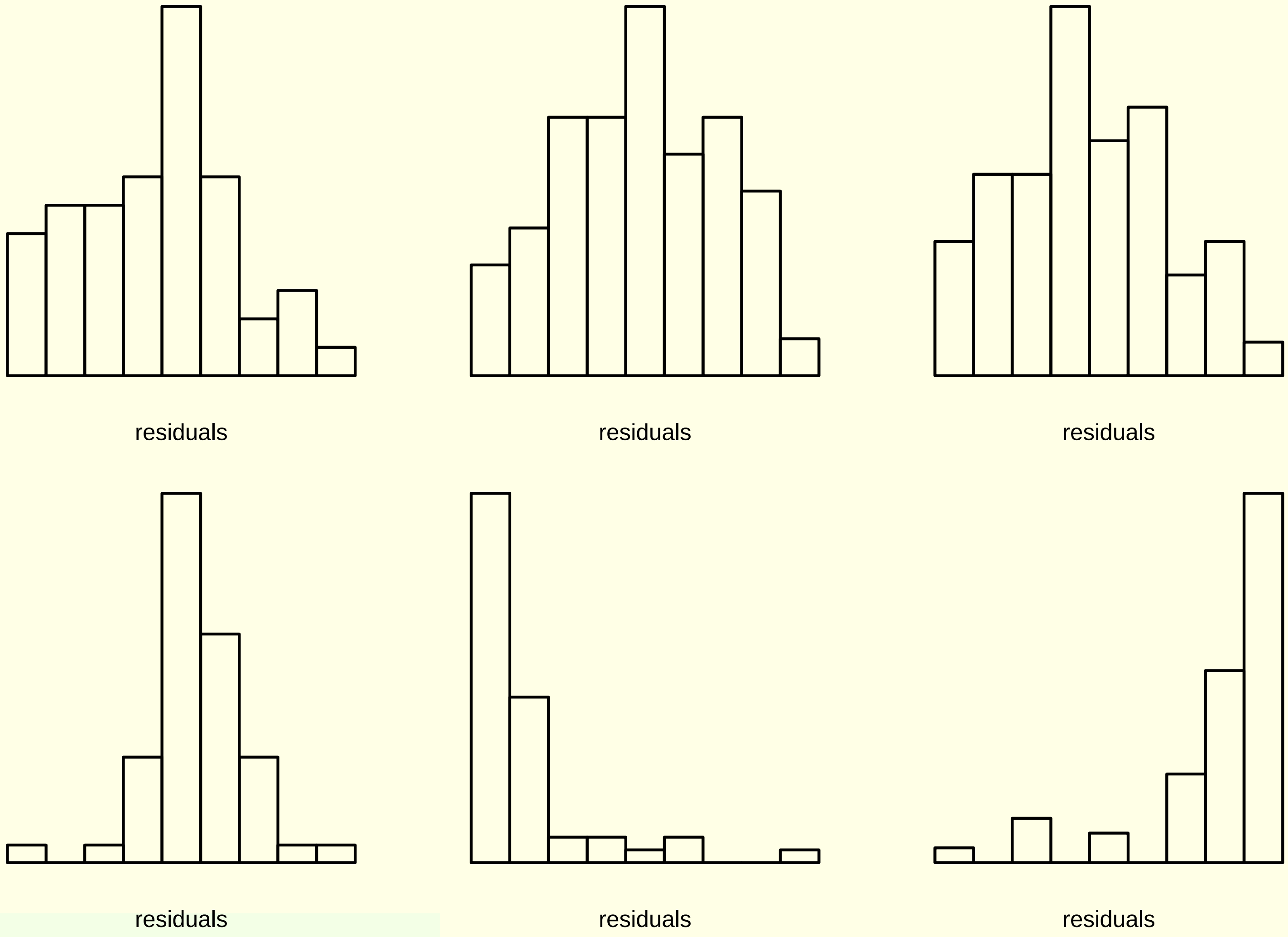
Residual plots which pass



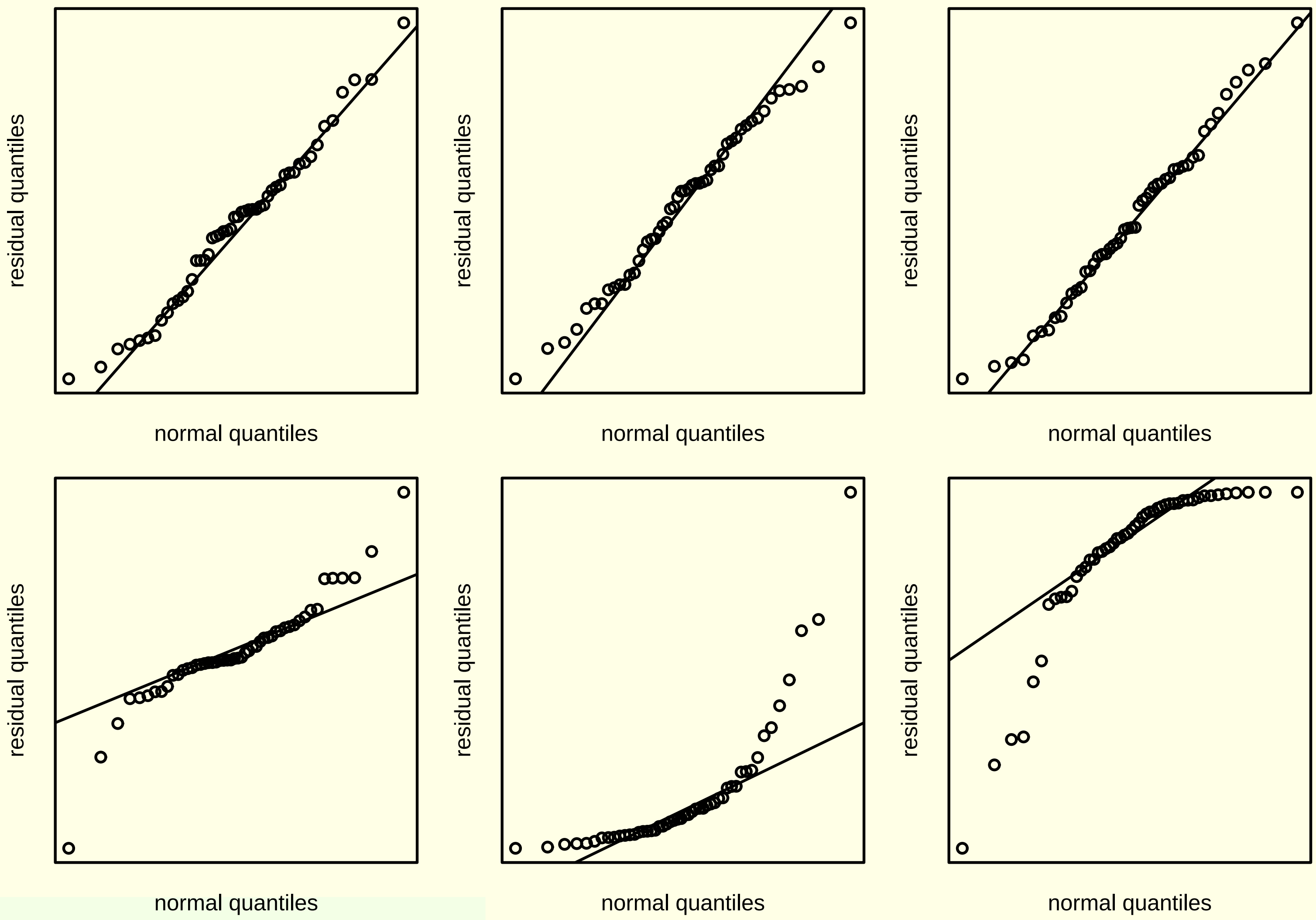
Residual plots which fail



Upper three pass, lower three fail



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 remedies to come in Chapters 3 and 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: **CARS2** data file—is weight or horsepower better for predicting cost?

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 [a]	0.945	0.927	2.7865

[a] Predictors: (Intercept), X.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(> |t|)
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

95% Confidence Interval

Model		Lower Bound	Upper Bound
1	(Intercept)	155.238	225.398
	X	22.712	58.888

[a] Response variable: Y.

Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between $Y =$ sale price (\$k) and $X =$ floor size (k sq. feet).

Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between $Y =$ sale price ($k) and $X =$ floor size (k sq. feet).
- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.

Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between $Y =$ sale price (\$k) and $X =$ floor size (k sq. feet).
- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- $X = 0$ does not make sense for this application, nor do we have data close to $X = 0$, so we cannot meaningfully interpret $\hat{b}_0 = 190.3$.

Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between $Y=$ sale price (\$k) and $X=$ floor size (k sq. feet).
- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- $X=0$ does not make sense for this application, nor do we have data close to $X=0$, so we cannot meaningfully interpret $\hat{b}_0 = 190.3$.
- Expect sale price to increase $\$4080$ when floor size increases 100 sq. feet, for 1683–2269 sq. feet homes (95% confident sale price increases between $\$2270$ and $\$5890$ when floor size increases 100 sq. feet).

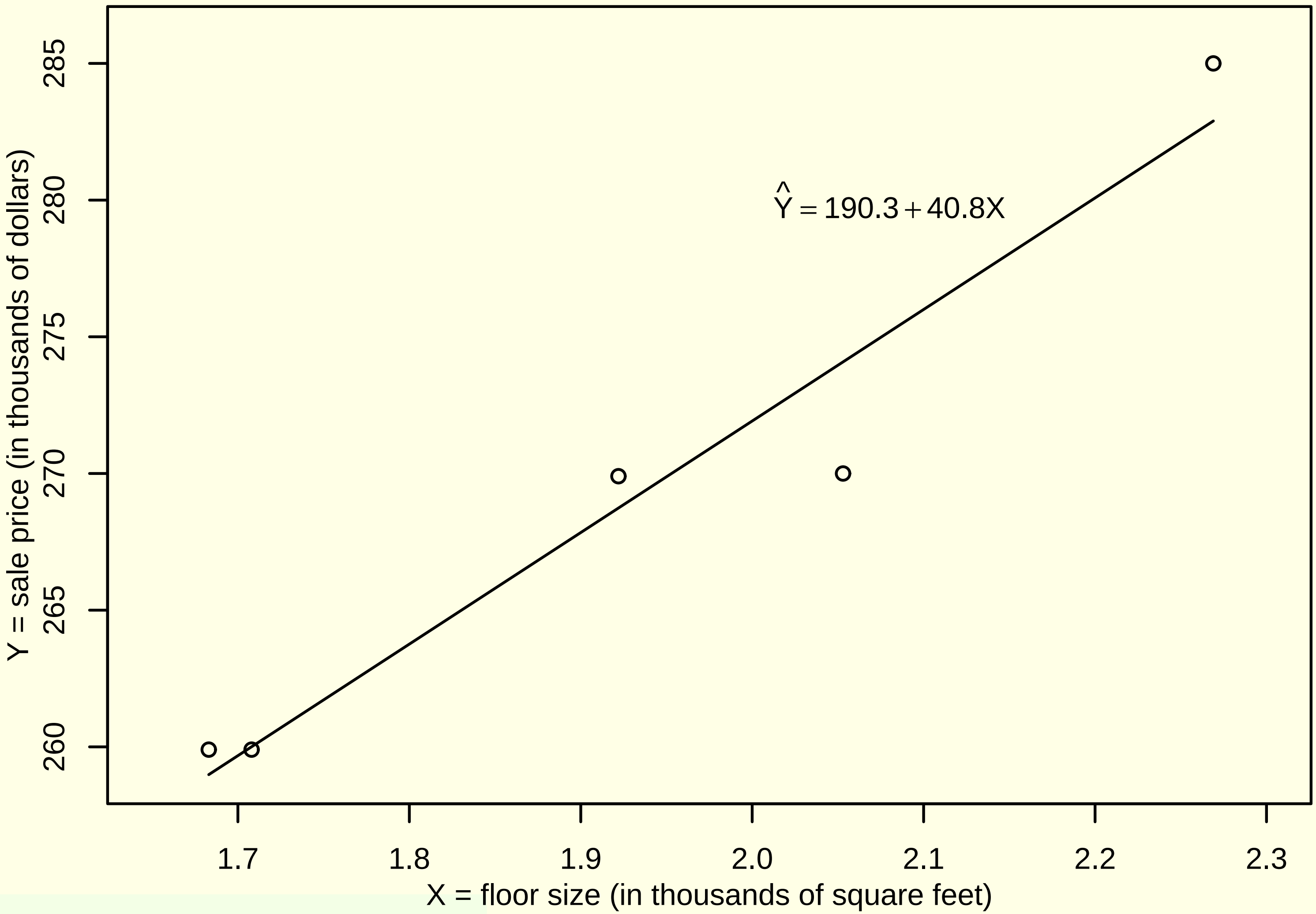
Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between $Y =$ sale price (\$k) and $X =$ floor size (k sq. feet).
- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- $X = 0$ does not make sense for this application, nor do we have data close to $X = 0$, so we cannot meaningfully interpret $\hat{b}_0 = 190.3$.
- Expect sale price to increase $\$4080$ when floor size increases 100 sq. feet, for 1683–2269 sq. feet homes (95% confident sale price increases between $\$2270$ and $\$5890$ when floor size increases 100 sq. feet).
- Can expect a prediction of an unobserved sale price from a particular floor size to be accurate to within approximately $\pm\$5570$ (with 95% confidence).

Interpreting model results

- We found a statistically significant straight-line relationship (at a 5% significance level) between $Y =$ sale price (\$k) and $X =$ floor size (k sq. feet).
- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- $X = 0$ does not make sense for this application, nor do we have data close to $X = 0$, so we cannot meaningfully interpret $\hat{b}_0 = 190.3$.
- Expect sale price to increase \$4080 when floor size increases 100 sq. feet, for 1683–2269 sq. feet homes (95% confident sale price increases between \$2270 and \$5890 when floor size increases 100 sq. feet).
- Can expect a prediction of an unobserved sale price from a particular floor size to be accurate to within approximately $\pm\$5570$ (with 95% confidence).
- 94.5% of the variation in sale price (about its mean) can be explained by a straight-line relationship between sale price and floor size.

Simple linear regression model.



Estimation and prediction

- Recall the confidence interval for a univariate population mean, $\mathrm{E}(Y)$:
  $m_Y \pm$ t-percentile$(s_Y/\sqrt{n})$.
- Also, a prediction interval for an individual univariate Y-value:
  $m_Y \pm$ t-percentile $\left(s_Y\sqrt{1+1/n}\right)$.

Estimation and prediction

- Recall the confidence interval for a univariate population mean, $\mathrm{E}(Y)$: $m_Y \pm$ t-percentile$(s_Y/\sqrt{n})$.
- Also, a prediction interval for an individual univariate Y-value: $m_Y \pm$ t-percentile $\left(s_Y\sqrt{1+1/n}\right)$.
- Similar distinction between confidence and prediction intervals for simple linear regression.
- Confidence interval for the population mean, $\mathrm{E}(Y)$, at a particular X-value is $\hat{Y} \pm$ t-percentile$(s_{\hat{Y}})$.
- Prediction interval for an individual Y-value at a particular X-value is $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$.

Estimation and prediction

- Recall the confidence interval for a univariate population mean, $\mathrm{E}(Y)$: $m_Y \pm$ t-percentile$(s_Y/\sqrt{n})$.
- Also, a prediction interval for an individual univariate Y-value: $m_Y \pm$ t-percentile $\left(s_Y\sqrt{1+1/n}\right)$.
- Similar distinction between confidence and prediction intervals for simple linear regression.
- Confidence interval for the population mean, $\mathrm{E}(Y)$, at a particular X-value is $\hat{Y} \pm$ t-percentile$(s_{\hat{Y}})$.
- Prediction interval for an individual Y-value at a particular X-value is $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$.
- Which should be wider? Is it harder to estimate a mean or predict an individual value?

Confidence interval for population mean, E(Y)

- Formula: $\hat{Y} \pm$ t-percentile$(s_{\hat{Y}})$
  where $s_{\hat{Y}} = s\sqrt{\frac{1}{n} + \frac{(X_p - m_X)^2}{\sum_{i=1}^{n}(X_i - m_X)^2}}$.
- Interval is narrower:
  - when n is large;
  - when X_p is close to its sample mean, m_X;
  - when the regression standard error, s, is small;
  - for lower levels of confidence.

Confidence interval for population mean, E(Y)

- Formula: $\hat{Y} \pm$ t-percentile$(s_{\hat{Y}})$
  where $s_{\hat{Y}} = s\sqrt{\frac{1}{n} + \frac{(X_p - m_X)^2}{\sum_{i=1}^{n}(X_i - m_X)^2}}$.
- Interval is narrower:
  - when n is large;
  - when X_p is close to its sample mean, m_X;
  - when the regression standard error, s, is small;
  - for lower levels of confidence.
- Example: for home prices–floor size dataset, the 95% confidence interval for $\mathrm{E}(Y)$ when $X = 2$ is $(267.7,\ 276.1)$.
- Interpretation: we're 95% confident that average sale price is between $\$267{,}700$ and $\$276{,}100$ for 2000 square foot homes.

Prediction interval for an individual Y-value

- Formula: $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$
  where $s_{\hat{Y}^*} = s\sqrt{1+\frac{1}{n}+\frac{(X_p - m_X)^2}{\sum_{i=1}^n (X_i - m_X)^2}}$.
- Interval is narrower:
  - when n is large;
  - when X_p is close to its sample mean, m_X;
  - when the regression standard error, s, is small;
  - for lower levels of confidence.

Prediction interval for an individual Y-value

- Formula: $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$
  where $s_{\hat{Y}^*} = s\sqrt{1 + \frac{1}{n} + \frac{(X_p - m_X)^2}{\sum_{i=1}^{n}(X_i - m_X)^2}}$.
- Interval is narrower:
  - when n is large;
  - when X_p is close to its sample mean, m_X;
  - 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.

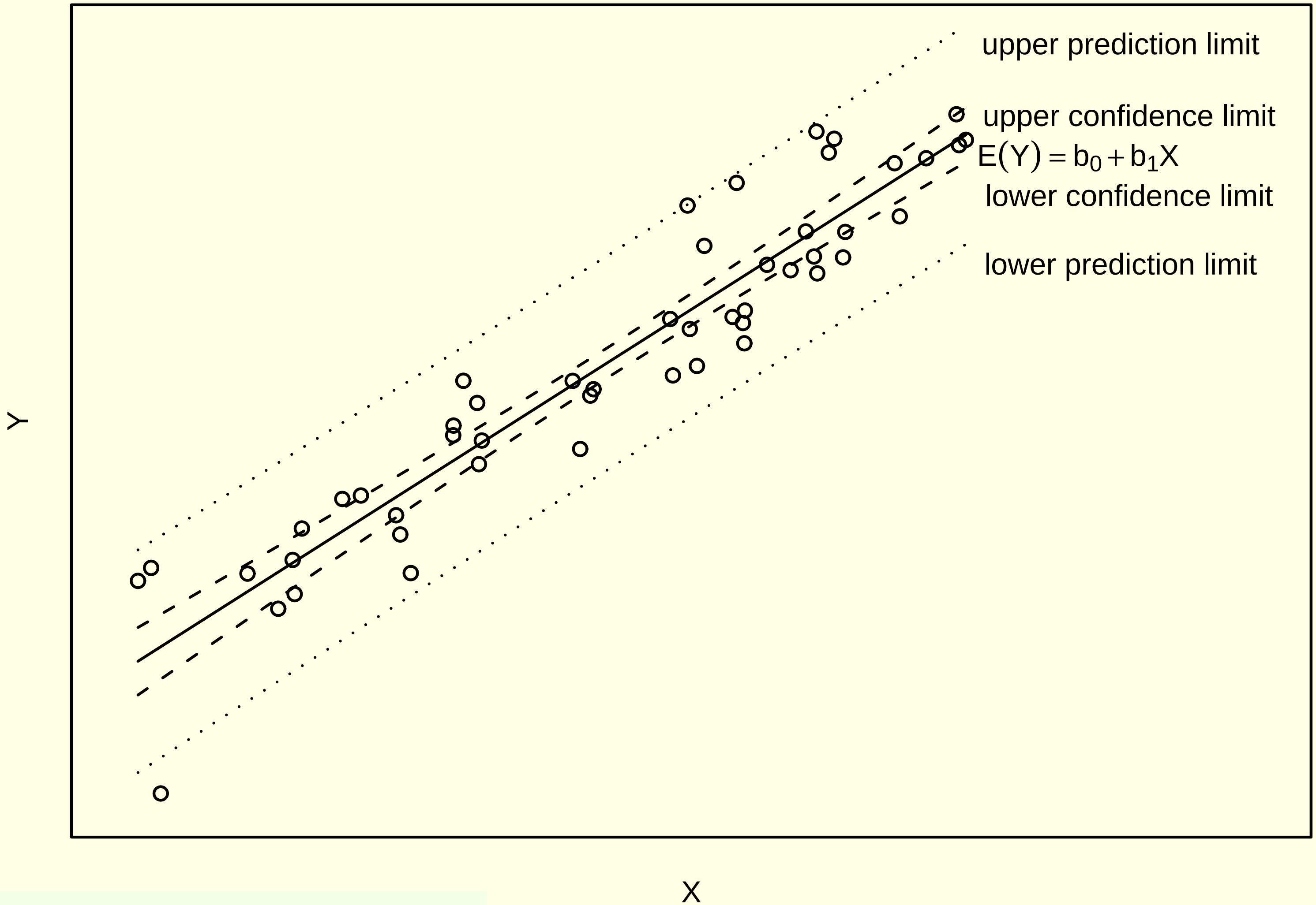
Prediction interval for an individual Y-value

- Formula: $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$
  where $s_{\hat{Y}^*} = s\sqrt{1 + \frac{1}{n} + \frac{(X_p - m_X)^2}{\sum_{i=1}^n (X_i - m_X)^2}}$.
- Interval is narrower:
  - when n is large;
  - when X_p is close to its sample mean, m_X;
  - 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: home prices–floor size dataset, the 95% prediction interval for Y^* at $X = 2$ is $(262.1, 281.7)$.
- Interpretation: we're 95% confident that the sale price for an individual 2000 square foot home is between $\$262{,}100$ and $\$281{,}700$.

Prediction interval for an individual Y-value

- Formula: $\hat{Y}^* \pm$ t-percentile$(s_{\hat{Y}^*})$ where $s_{\hat{Y}^*} = s\sqrt{1 + \frac{1}{n} + \frac{(X_p - m_X)^2}{\sum_{i=1}^n (X_i - m_X)^2}}$.
- Interval is narrower:
  - when n is large;
  - when X_p is close to its sample mean, m_X;
  - 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: home prices–floor size dataset, the 95% prediction interval for Y^* at $X = 2$ is $(262.1, 281.7)$.
- Interpretation: we're 95% confident that the sale price for an individual 2000 square foot home is between $\$262{,}100$ and $\$281{,}700$.
- What is a 95% prediction interval for large n?

Compare widths of confidence and prediction intervals.



Steps in a simple linear regression analysis

- Formulate model.

Steps in a simple linear regression analysis

- Formulate model.
- Construct a scatterplot of Y versus X.

Steps in a simple linear regression analysis

- Formulate model.
- Construct a scatterplot of Y versus X.
- Estimate model using least squares.

Steps in a simple linear regression analysis

- Formulate model.
- Construct a scatterplot of Y versus X.
- Estimate model using least squares.
- Evaluate model:
  - Regression standard error, s;
  - Coefficient of determination, R^2;
  - Population slope, b_1.

Steps in a simple linear regression analysis

- Formulate model.
- Construct a scatterplot of Y versus X.
- Estimate model using least squares.
- Evaluate model:
  - Regression standard error, s;
  - Coefficient of determination, R^2;
  - Population slope, b_1.
- Check model assumptions.

Steps in a simple linear regression analysis

- Formulate model.
- Construct a scatterplot of Y versus X.
- Estimate model using least squares.
- Evaluate model:
  - Regression standard error, s;
  - Coefficient of determination, R^2;
  - Population slope, b_1.
- Check model assumptions.
- Interpret model.

Steps in a simple linear regression analysis

- Formulate model.
- Construct a scatterplot of Y versus X.
- Estimate model using least squares.
- Evaluate model:
  - Regression standard error, s;
  - Coefficient of determination, R^2;
  - Population slope, b_1.
- Check model assumptions.
- Interpret model.
- Estimate $\mathrm{E}(Y)$ and predict Y.