Assumptions For Regression Inferences

1. Population Regression Line (Assumption I): There is a straight line, \( y = \beta_0 + \beta_1 x \) such that for each \( x \)-value, the mean of the corresponding population of \( y \)-values lies on that straight line.
2. **Equal Standard Deviations (Assumption II):** The standard deviation, $\sigma$, of the population of $y$-values corresponding to a particular $x$-value is the same, regardless of the $x$-value.

3. **Normality (Assumption III):** For each $x$-value, the corresponding population of $y$-values is normally distributed.

**FIGURE 13.4** Graphical portrayal of assumptions for regression inferences for Nissan Z illustration
**Standard Error Of The Estimate**

The standard error of the estimate is defined by:

\[ s_e = \sqrt{\text{SSE} / (n - 2)} \]

It provides us with an estimate for the population standard deviation. The \( s_e \) indicates how far the observed y-values are from the predicted y-values, on average.

**Residual Analysis For the Regression Model**

If the assumptions for regression inferences are met, then the following two conditions should hold.

1. A plot of the residuals against the x-values should fall roughly in a horizontal band centered and symmetric about the x-axis. If a pattern is indicated you probably need to use some other analytical method that simple linear regression. For example, the following graph shows that a quadratic relationship exists in the residuals.
2. A normal probability plot of the residuals should be roughly linear. For example, the following graph shows that the normality assumption is violated.
Inferences for the Slope of the Population Regression Line

H₀: β₁ = 0
Hₐ: β₁ ≠ 0

The test statistic has a t-distribution with df=(n - 2).

\[ t = \frac{(b_1 - \beta_1)}{(s_e / \sqrt{S_{xx}})} \]

If the value of the test statistic falls in the rejection region, then reject the null; otherwise do not reject the null.

Confidence Intervals for the Slope

The endpoints of the confidence interval for β₁ are:

\[ b_1 \pm t_{\alpha/2} \left( \frac{s_e}{\sqrt{S_{xx}}} \right) \quad \text{with df = n-2} \]

Confidence Intervals for Means in Regression – (skip Sta308)

\[ \hat{y}_p \pm t_{\alpha/2}S_e \sqrt{\frac{1}{n} + \frac{x_p - \overline{x}}{\sum x^2 / n}} \quad \text{with df = (n -2)} \]

Confidence Intervals for a Population y-value given an x-value (skip Sta308)

\[ \hat{y}_p \pm t_{\alpha/2}S_e \sqrt{1 + \frac{1}{n} + \frac{x_p - \overline{x}}{\sum x^2 / n}} \quad \text{with df = (n -2)} \]
Regression Analysis: Inferential & Correlation

Regression Example  << Regrsam.xls >>

<table>
<thead>
<tr>
<th>Age (x)</th>
<th>Price (y)</th>
<th>xy</th>
<th>Sqr(x)</th>
<th>Sqr(y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>125</td>
<td>750</td>
<td>36</td>
<td>15625</td>
</tr>
<tr>
<td>6</td>
<td>115</td>
<td>690</td>
<td>36</td>
<td>13225</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
<td>780</td>
<td>36</td>
<td>16900</td>
</tr>
<tr>
<td>2</td>
<td>260</td>
<td>520</td>
<td>4</td>
<td>67600</td>
</tr>
<tr>
<td>2</td>
<td>219</td>
<td>438</td>
<td>4</td>
<td>47961</td>
</tr>
<tr>
<td>5</td>
<td>150</td>
<td>750</td>
<td>25</td>
<td>22500</td>
</tr>
<tr>
<td>4</td>
<td>190</td>
<td>760</td>
<td>16</td>
<td>36100</td>
</tr>
<tr>
<td>5</td>
<td>163</td>
<td>815</td>
<td>25</td>
<td>26569</td>
</tr>
<tr>
<td>1</td>
<td>260</td>
<td>260</td>
<td>1</td>
<td>67600</td>
</tr>
<tr>
<td>4</td>
<td>160</td>
<td>640</td>
<td>16</td>
<td>25600</td>
</tr>
</tbody>
</table>

TOTALS  41  1772  6403  199  339680

\[ Syy = 339680 - \frac{Sqr(1772)}{10} = 25681.60 \]
\[ Sxx = 199 - \frac{Sqr(41)}{10} = 30.90 \]
\[ Sxy = 6403 - (41)(1772) / 10 = -862.20 \]
\[ B1 = \frac{Sxy}{Sxx} = -27.90 \]
\[ Bo = \frac{1}{11} (1772 - b1(41)) = 265.09 \]

\[ SST = Syy = 25681.600 \]
\[ SSR = \frac{(Sxy*Sxy)}{Sxx} = 24057.891 \]
\[ SSE = SST - SSR = 1623.709 \]
\[ Se = \sqrt{\frac{SSE}{(n-2)}} = 14.247 \]
\[ r\text{-square} = \frac{(1-(SSE / SST))*100}{93.68\%} \]
\[ r = \sqrt{r\text{-square}} = 0.97 \]
Hypothesis Testing

Ho: B1 = 0  
Ha: B1 <> 0

t = b1 / (Se / Sqrt(Sxx)) = -10.887

Critical t at 95% and df=8 = 2.306

Since calculated value is in the rejection region we reject the null hypothesis.

There is enough evidence to conclude that the age of corvettes is useful for predicting price of the corvettes.

Confidence Intervals at 95%

-27.9 - 2.306 * (14.247 / Sqrt(30.90)) = -33.81
CI = (-21.99, -33.81)
Multiple Regression

<< only via excel – output analysis >>

Aside: The F distribution

- Not symmetric
- Does not have zero at its center

Using the F-Table

- Need a significance level, numerator degrees of freedom \((n_1 - 1)\), and denominator degrees of freedom \((n_2 - 1)\)
- To find the left tailed critical value we can use the reciprocal of the right-tailed value with the numbers of degrees of freedom reversed.
- For example: if \(n_1 = 10\), \(n_2 = 7\), \(\alpha = 0.05\) the right tail value is 5.5234 (from table). The left tail value is \((1/4.317 = 0.2315)\). That is, 4.317 is the critical value for 6,9 degrees of freedom.