HW # 4 Solutions

(1) 8.4

a. The estimated model is

\[ \hat{Y} = 82.9 - 1.18x + 0.0148x^2 \]

\[ R^2 = 0.763 \]

Considering the plot to the right, and the fairly large \( R^2 \), yes, this model appears to fit the data well.

b. \( H_0: \beta_1 = \beta_{11} = 0; \ Ha: \) not both = 0.

\[ SSR = 11830.6, \text{df} = 2, \text{MSR} = 5915.31. \]

\[ SSE = 3671.3, \text{df} = 57, \text{MSE} = 64.41. \]

\[ F^* = 91.84. \quad P \ll 0.0001. \] Significant at \( \alpha = 0.05 \), so reject \( H_0 \). (The \( F \) critical value is 3.15884.)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>11830.6</td>
<td>5915.3</td>
<td>91.84</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>57</td>
<td>3671.3</td>
<td>64.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>15501.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e. \( H_0: \beta_{11} = 0; \ Ha: \beta_{11} \neq 0. \)

By \( t \) test:

\[ s\{b_{11}\} = 0.008357. \quad t^* = b_{11} / s\{b_{11}\} = 0.014840 / 0.008357 = 1.7759, \quad P\text{-value} = 0.081 \]

Conclusion: do not reject \( H_0 \). (The \( t \) critical value is 2.00247.)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>c_age_sqd</td>
<td>0.014840</td>
<td>0.008357</td>
<td>1.78</td>
<td>0.081</td>
</tr>
</tbody>
</table>

Alternatively and equivalently, by added-SS \( F \) test:

\[ SSR(x^2 | x) = 203.1. \quad SSE(x, x^2) = 3671.31, \text{MSE} = 64.4. \quad F^* = 203.1 / 64.4 = 3.1533, \]

\[ P\text{-value} = 0.081. \] Conclusion: do not reject \( H_0 \). (The \( F \) critical value is 4.00987.)

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>c_age_sqd</td>
<td>1</td>
<td>203.1</td>
</tr>
</tbody>
</table>

g. \( r_{X, X^2} = 0.9961, \quad r_{X, X^2} = -0.0384. \)

Yes, centering the age variable before squaring it greatly reduced the correlation between the linear and squared terms.
(2) 8.5

a. There is nothing remarkable about these residual plots; the model appears to be appropriate for these data.

b. \( H_0: E\{Y\} = \beta_0 + \beta_1 x + \beta_{11} x^2 \) (i.e. the model for the mean \( Y \) is correct),
\( H_a: E\{Y\} \neq \beta_0 + \beta_1 x + \beta_{11} x^2 \) (i.e. the model for the mean \( Y \) is not correct)

\[
\begin{array}{l|cccc}
\text{Source} & \text{DF} & \text{SS} & \text{MS} & \text{F} & \text{P} \\
\hline
\text{Residual Error} & 57 & 3671.3 & 64.4 & & \\
\text{Lack of Fit} & 29 & 1821.6 & 62.8 & 0.95 & 0.554 \\
\text{Pure Error} & 28 & 1849.7 & 66.1 & & \\
\end{array}
\]

With \( P > \alpha \), we do not reject \( H_0 \). (The \( F \) critical value is 1.875.)

The assumptions are the same as in ANOVA, all of which are shared with regression: observations are independent, the distribution of \( Y \) for given \( X \) is normally distributed, and the variance of the \( Y \)s is the same for all \( X \).

c. The estimated model is
\[
\hat{Y} = 82.927 - 1.26789 \cdot x + 0.01504 \cdot x^2 + 0.000337 \cdot x^3
\]
\( H_0: \beta_{111} = 0; \ H_a: \beta_{111} \neq 0. \)

By \( t \) test (the \( t \) critical value is 2.003):

\[
\begin{array}{lcccc}
\text{Predictor} & \text{Coef} & \text{SE Coef} & \text{T} & \text{P} \\
\text{c_age_cubed} & 0.0003369 & 0.0009327 & 0.36 & 0.719 \\
\end{array}
\]
By added-SS and $F$ test:

$$F^* = \frac{8.5}{65.4} = 0.13, \ P\text{-value} = 0.719.$$ (The $F$ critical value is 4.013.)

We do not reject $H_0$. This conclusion is consistent with the LOF test for the quadratic model in part b, which concluded there was no statistically significant evidence of higher-order effects.

(3) a. No, the relationship does not appear to be the same for all regions: The relationships in Regions 1–3 are quite similar, but in Region 4, although the slope is similar to that in the other regions, the number of beds is smaller for a given population.

b. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_{i1} + \beta_3 Z_{i2} + \beta_4 Z_{i3} + \beta_5 X_i Z_{i1} + \beta_6 X_i Z_{i2} + \beta_7 X_i Z_{i3} + \epsilon_i$.

$[Z_{i1} = 1 \text{ if region} = 1, 0 \text{ otherwise}; \text{and similarly for} Z_{i2} \text{ (region 2) and} Z_{i3} \text{ (region 3).}]$

$H_0: \beta_2 = \beta_3 = \ldots = \beta_7 = 0; \ H_a: \text{not all} = 0.$

**indicator-variable regression**, using 1/0 coding:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>7</td>
<td>60.4206</td>
<td>8.6315</td>
<td>162.62</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>432</td>
<td>22.9302</td>
<td>0.0531</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>439</td>
<td>83.3508</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Diagram:**

- **Region 1**
- **Region 2**
- **Region 3**
- **Region 4**
multiple-partial $F$ test using these sequential sums-of-squares:

$$SS(Z_1, Z_2, Z_3, XZ_1, XZ_2, XZ_3 | X) = 0.0107 + ... + 0.0035 = 1.9825;$$
$$df = 6; \quad MS = 0.3304; \quad F^* = 0.3304 / 0.0531 = 6.2225; \quad P < 0.0001. \quad \text{Reject } H_0.$$

**GLM:**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>lgPop</td>
<td>1</td>
<td>58.4381</td>
<td>59.4937</td>
<td>59.4937</td>
<td>1120.85</td>
<td>0.000</td>
</tr>
<tr>
<td>Region</td>
<td>3</td>
<td>1.9206</td>
<td>0.0456</td>
<td>0.0152</td>
<td>0.29</td>
<td>0.835</td>
</tr>
<tr>
<td>Region*lgPop</td>
<td>3</td>
<td>0.0619</td>
<td>0.0619</td>
<td>0.0206</td>
<td>0.39</td>
<td>0.761</td>
</tr>
<tr>
<td>Error</td>
<td>432</td>
<td>22.9302</td>
<td>22.9302</td>
<td>0.0531</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>439</td>
<td>83.3508</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

multiple-partial $F$ test using these sequential sums-of-squares:

$$SS(Z_1, Z_2, Z_3, XZ_1, XZ_2, XZ_3 | X) = 1.9206 + 0.0619 = 1.9825;$$
$$df = 3 + 3 = 6; \quad MS = 0.3304; \quad F^* = 0.3304 / 0.0531 = 6.2225; \quad P < 0.0001. \quad \text{Reject } H_0.$$

**Conclusion:** The four regression relationship are not all the same. (How they differ is not determined from these tests.)

c. $H_0: \beta_5 = \beta_6 = \beta_7 = 0; \quad H_a: \text{not all = 0}$

multiple-partial $F$ test using the indicator-variable regression sequential SSes in part b :

$$SS(XZ_1, XZ_2, XZ_3 | X, Z_1, Z_2, Z_3) = 0.0488 + 0.0096 + 0.0035 = 0.0619;$$
$$df = 3; \quad MS = 0.0.0206; \quad F^* = 0.0206 / 0.0531 = 0.389; \quad P = 0.76. \quad \text{Do not reject } H_0.$$

**GLM** test is directly in the GLM output above, as the Region*lgPop term.

**Conclusion:** The four slopes are not statistically significantly different: the regression relationships are parallel.

d. Yes, the assumption seems reasonable. There is somewhat more spread in Regions 2 and 3 than in Regions 1 and 4, but not enough so to invalidate the analysis. (This comparison could be done more quantitatively by calculating the standard deviations of the residuals for each region; these are 0.1596, 0.2537, 0.2672, and 0.1873 respectively, within the 2x guideline for ANOVA.)
e. The plots are shown on the next page. There are three mildly noteworthy features:

- Even with the log transformation there are a few unusually large counties (especially Los Angeles and Cook [Chicago], also Harris [Houston], San Diego, Orange, and Kings), which will have substantial leverage.
- There appears to be less variability among the larger counties, especially in Region 2, though this appearance is at least partly due to there being many more observations with smaller than larger populations.
- Fairfax Co., VA, is a particularly low outlier in Region 3.

f. The estimated functions for Regions 1–3 are almost identical, while that for Region 4 is somewhat below the others. In the full model (with region x log(population) interaction), the slope for Region 1 is slightly steeper than that for the other regions.

If including such a graph in a manuscript I would use the one with parallel regressions, as that was the conclusion from the analysis, but an argument could be made for presenting the full model instead, so that readers could see for themselves how the slopes vary.