2.17 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

a. Plot house price against house size in a scatter diagram.

b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

d. Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.

e. For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.

f. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?

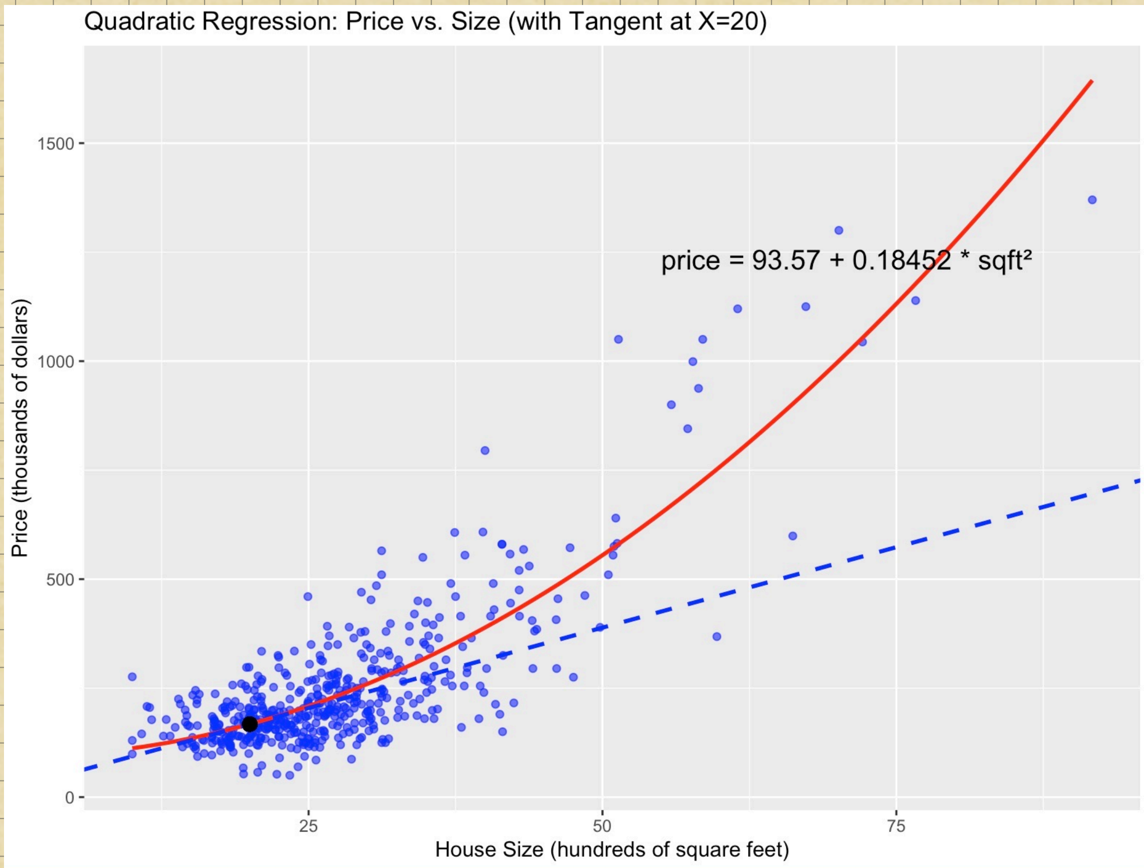
g. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a "better-fitting" model?

a,

b.

c.

d.



$$\frac{d\,PRICE}{d\,SQFT} = 2a_2 SQFT$$

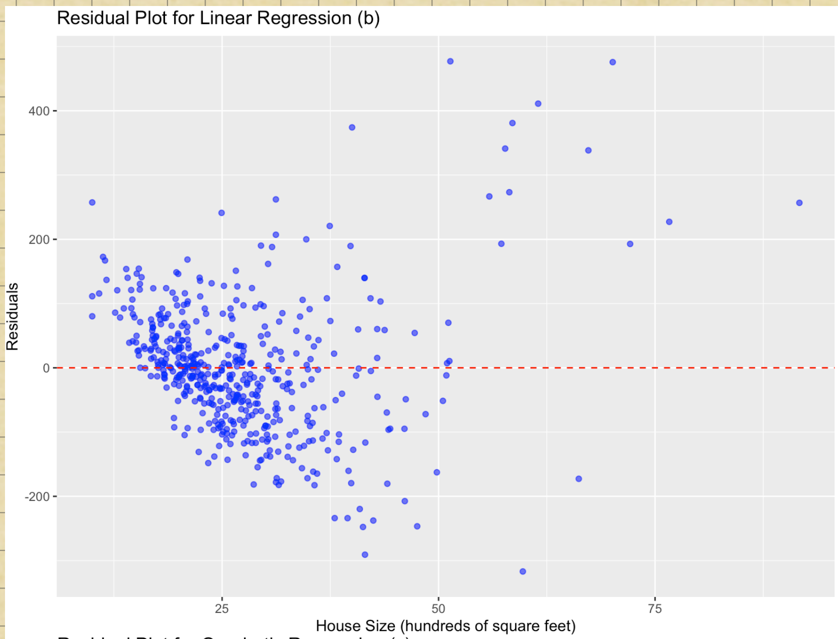
$\Rightarrow$ Marginal Effect $= 2 \times 0.18452 \times 20 = 7.3808$

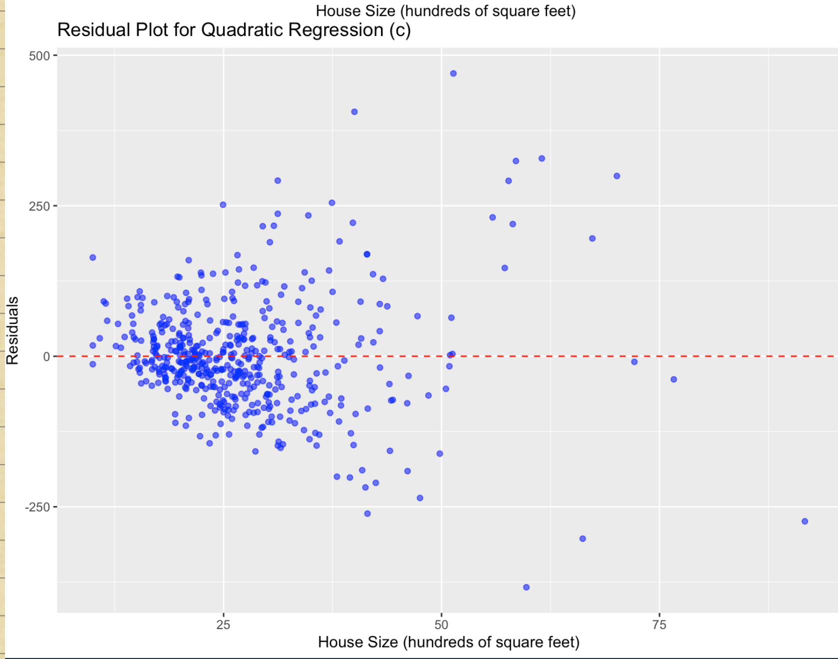
e. $$\varepsilon = \frac{d\,PRICE}{d\,SQFT} \times \frac{SQFT}{PRICE} = 2a_2 SQFT \times \frac{SQFT}{PRICE}$$

$E(PRICE \mid SQFT = 20) = 167.378$

$\varepsilon = \underline{0.8819}$ #

f.





Heteroscedasticity and Its Implications

We observe that residuals become more spread out as sqft increases, it means that the variance of the errors is not constant. This is a violation of the OLS assumption of homoscedasticity. Heteroscedasticity often indicates that:

The model does not properly capture the relationship between sqft and price, meaning a transformation might be necessary.

Prediction errors are larger for larger homes, which may suggest that a simple linear model is insufficient.

g.

```
> cat("Linear  SSE:", SSE_linear, "\n")
Linear  SSE: 5262847
> cat("Quadratic SSE:", SSE_quad, "\n")
Quadratic SSE: 4222356
```

The Quadratic model has lower SSE

A lower **Sum of Squared Residuals (SSE)** indicates:

- **Smaller prediction errors** → The predicted values are closer to the actual values.
- **Better explanation of data variability** → The model captures the variations in the data more effectively.

However, **SSE alone should not be the sole criterion for model selection**, because:

- More complex models (such as quadratic or polynomial regression) tend to **automatically reduce SSE**, but they may also lead to **overfitting**.
- **It is important to also consider Adjusted R^2, AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion)** to balance model fit and complexity.

📌 **Summary:**

- **Comparing SSE helps identify the model with a better fit.**
- **Quadratic regression typically has a lower SSE, but overfitting should be considered.**
- **The best model should strike a balance between goodness of fit and model simplicity.** 🚀📊