The image you provided is an **influence plot** (or **bubble plot**) used in **regression diagnostics**. Here’s what it tells us:

**Key Observations:**

1. **X-axis (hat values)**: Represents **leverage**, indicating how much influence a data point has on the regression model.
   * Higher **hat values** mean the point is an outlier in terms of predictor variables.
2. **Y-axis (standardized residuals)**: Shows how far a point’s actual value deviates from the predicted value.
   * Large positive or negative **residuals** indicate that the model is making large errors for those observations.
3. **Bubble size (Cook’s Distance)**: Represents **influence** (a combination of leverage and residual size).
   * **Larger bubbles** indicate that removing those data points would significantly alter the regression model.

**Interpretation of This Plot:**

* There are **four influential points** (large bubbles).
* These points have either:
  + **High leverage** (far right on the x-axis).
  + **High residuals** (far from 0 on the y-axis).
  + **Both** (indicating strong influence on the regression).
* **Data points on the upper or lower extremes of the y-axis** (±2.5) are potential **outliers**.
* **Data points with high hat values and large Cook’s distance** are highly **influential observations**, meaning they might be distorting the regression model.

**house\_98105 <- house\_df[house\_df$zipcode == 98105,]**

**lm\_98105 <- lm(price ~ sqft\_living + sqft\_lot + bathrooms +**

**bedrooms + grade, data=house\_98105)**

**house\_98105 ## 182 rows**

**std\_resid <- rstandard(lm\_98105)**

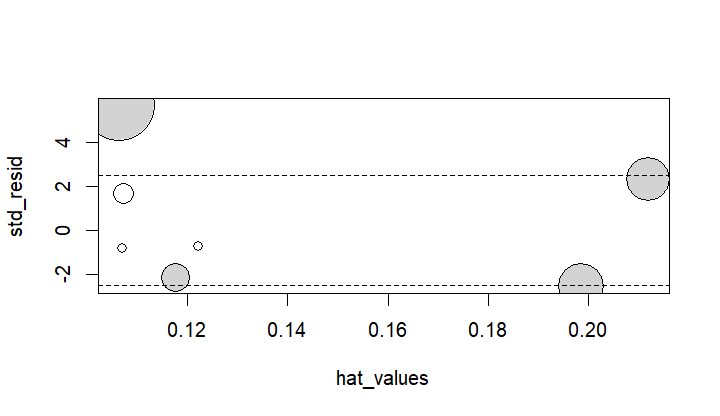
**cooks\_D <- cooks.distance(lm\_98105)**

**hat\_values <- hatvalues(lm\_98105)**

**plot(subset(hat\_values, cooks\_D > 0.08), subset(std\_resid, cooks\_D > 0.08),**

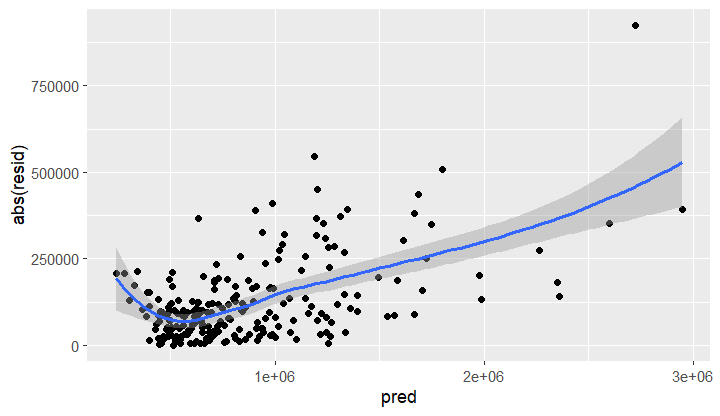
**xlab='hat\_values', ylab='std\_resid',**

**cex=10\*sqrt(subset(cooks\_D, cooks\_D > 0.08)), pch=16, col='lightgrey')**

****

**df <- data.frame(resid = residuals(lm\_98105), pred = predict(lm\_98105))**

**ggplot(df, aes(pred, abs(resid))) + geom\_point() + geom\_smooth()**

****

**Interpretation of the Residual Plot**

The provided graph is a **heteroskedasticity check** for the regression model lm\_98105. It plots **absolute residuals** (abs(resid)) against **predicted values** (pred).

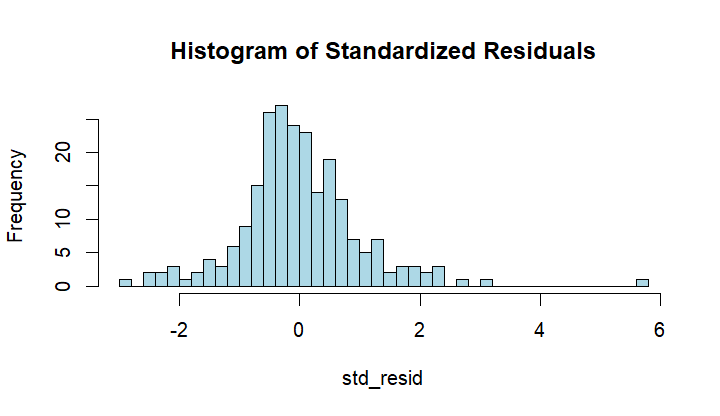
**Key Observations:**

1. **X-axis (predicted values)**: Represents the **fitted values** (predictions from the regression model).
2. **Y-axis (absolute residuals)**: Shows the **absolute magnitude of errors** (how far actual values deviate from predicted ones).
3. **Scatter points (black dots)**: Each point represents a single observation’s residual.
4. **Blue line (LOESS smoother)**: Indicates the **trend** in the residuals.
5. **Shaded region (confidence band)**: Represents **uncertainty** around the trend.

**Analysis of the Residual Pattern:**

* **Heteroskedasticity detected**: The residual variance **increases with predicted values**.
  + Lower predicted values have **lower and more consistent residuals**.
  + Higher predicted values show **larger and more spread-out residuals**.
* **Curved trend**: The blue line is **not flat**, which suggests:
  + The model might be missing an important variable.
  + Non-linearity in the data (potential need for transformations).
* **Outliers in high predictions**: Some extreme points have **very large residuals**, further confirming variance instability.

**hist(std\_resid, breaks=50, col="lightblue", main="Histogram of Standardized Residuals")**

****

**Key Observations:**

1. **Centering around Zero**:
   * The histogram is **centered around 0**, which is expected for residuals.
   * This suggests that the model does not have a strong systematic bias.
2. **Skewness and Long Tails**:
   * The right tail (positive residuals) appears **longer than the left tail**, indicating **right-skewness**.
   * Some **large positive residuals** suggest the presence of **outliers** where the model underestimates the actual values.
3. **Non-Normal Distribution**:
   * Ideally, residuals should follow a **bell-shaped normal distribution**.
   * Here, the distribution appears **slightly asymmetric** and **right-skewed**.
   * **Potential issues**:
     + Missing predictors in the model.
     + Heteroskedasticity (as seen in the previous residual plot).
     + Outliers impacting the fit.

**Implications of Non-Normal Residuals:**

* **If the goal is prediction**, this may not be a big issue.
* **If statistical inference is important (p-values, confidence intervals, hypothesis testing)**:
  + **Violating normality** affects the validity of these statistical measures.
  + In this case, a transformation (e.g., log transformation of price) might help.

**Interpretation of the Residual Plot and Durbin-Watson Test**

This graph is a **residual plot** for the regression model lm\_98105, created using ggplot2 in R. It shows the **absolute residuals** (abs(resid)) against **predicted values** (pred), with a LOESS smoother.

**Key Observations:**

1. **X-axis (predicted values)**: Represents the **fitted values** (model predictions).
2. **Y-axis (absolute residuals)**: Shows the **magnitude of residual errors**.
3. **Scatter points (black dots)**: Each dot represents an observation’s residual.
4. **Blue line (LOESS smoother)**: Shows the trend in the residuals.
5. **Shaded region**: Represents **confidence intervals** around the LOESS fit.

**Analysis:**

**1. Heteroskedasticity is Present**

* The **variance of residuals increases as predictions increase**.
* This suggests **heteroskedasticity**, meaning the model **does not maintain a constant error variance**.
* **Implication**: Predictions for **higher-priced houses** have **larger errors**.

**2. Possible Model Issues**

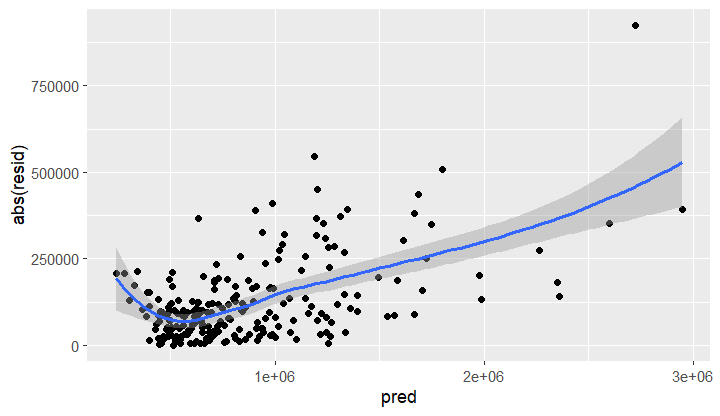
* The residuals show a **non-random pattern**, suggesting:
  + **Missing predictor variables** (e.g., location, home renovations).
  + **Non-linearity** in relationships (e.g., price may not increase linearly with sqft\_living).
* The **right tail (higher predicted values) has higher variance**, suggesting the model underestimates high-value properties.

**Durbin-Watson Test (dwtest(lm\_98105))**

* The **Durbin-Watson test** checks for **autocorrelation in residuals**.
* **If DW statistic ≈ 2** → Residuals are independent (good).
* **If DW < 2** → **Positive autocorrelation** (previous residuals influence the next).
* **If DW > 2** → **Negative autocorrelation** (residuals alternate in sign).
* **Low p-value (< 0.05)** indicates significant autocorrelation.

**If Autocorrelation Exists:**

* This suggests a pattern in errors, meaning **predictors may be missing an important time-related or spatial factor**.
* **Solution:** Consider **time-series regression or spatial modeling**.

****

[**https://chatgpt.com/c/67bc92f2-07cc-8004-b16a-1880d34c7e75**](https://chatgpt.com/c/67bc92f2-07cc-8004-b16a-1880d34c7e75)