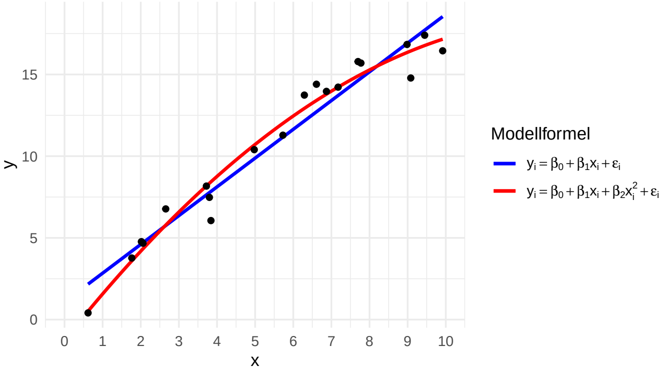
Was soll der Ramsey RESET Test prüfen?

Regression **E**quation **S**pecification **E**rror **T**est von James **Ramsey** (1969)

Was soll der Ramsey RESET Test prüfen?

Regression **E**quation **S**pecification **E**rror **T**est von James **Ramsey** (1969)

Ist die Modellgleichung im Regressionsmodell fehlspezifiziert?



Welches Modell ist richtig?

Wir können das quadratische Modell schätzen und mit dem t-Test überprüfen, ob β_2 signifikant (von Null verschieden) ist:

```
formula <- y ~ x + I(x^2)
quadratic_model <- lm(formula, data)
```

Welches Modell ist richtig?

Wir können das quadratische Modell schätzen und mit dem t-Test überprüfen, ob β_2 signifikant (von Null verschieden) ist:

```r
formula <- y ~ x + I(x^2)
quadratic_model <- lm(formula, data)
```

```r
summary(quadratic_model)$coefficients
```

```
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)    -1.27      0.816    -1.6 0.13864997
## x               2.95      0.345     8.5 0.00000015
## I(x^2)         -0.11      0.031    -3.5 0.00258782
```

Welches Modell ist richtig?

Wir können das quadratische Modell schätzen und mit dem t-Test überprüfen, ob β_2 signifikant (von Null verschieden) ist:

```r
formula <- y ~ x + I(x^2)
quadratic_model <- lm(formula, data)
```

```r
summary(quadratic_model)$coefficients
```

```
##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)    -1.27      0.816    -1.6 0.13864997
## x               2.95      0.345     8.5 0.00000015
## I(x^2)         -0.11      0.031    -3.5 0.00258782
```

Hier ist der p-value für $H_0 : \beta_2 = 0$ ziemlich klein, also wird H_0 verworfen. Wir würden uns also für das quadratische Modell entscheiden.

Funktioniert dieses Vorgehen auch bei mehreren Regressoren?

Bei $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ gegen $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ hat das gut funktioniert.

Und wenn wir mehr als nur einen Regressor haben?

Funktioniert dieses Vorgehen auch bei mehreren Regressoren?

Bei $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ gegen $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ hat das gut funktioniert.

Und wenn wir mehr als nur einen Regressor haben?

$$\text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i + \epsilon_i$$

Funktioniert dieses Vorgehen auch bei mehreren Regressoren?

Bei $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ gegen $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ hat das gut funktioniert.

Und wenn wir mehr als nur einen Regressor haben?

$$\text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i + \epsilon_i$$

Hier gibt es viel mehr Möglichkeiten:

- Quadratische Terme: Wohnfläche_i^2, $\text{Schlafzimmeranzahl}_i^2$, Baujahr_i^2
- Interaktionsterme: $\text{Wohnfläche}_i \cdot \text{Schlafzimmeranzahl}_i, \ldots$
- Kubische Terme: $\text{Wohnfläche}_i^3, \ldots$

Für jede Möglichkeit einen t-Test durchzuführen ist zu aufwendig.

Was ist nun Ramsey's Idee?

1. Schätze das Ausgangsmodell:

$$\text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i + \epsilon_i$$

Was ist nun Ramsey's Idee?

1. Schätze das Ausgangsmodell:

$$\text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i + \epsilon_i$$

2. Berechne die Prognosen:

$$\color{red}{\widehat{\text{Hauspreis}}_i} = \hat{\beta}_0 + \hat{\beta}_1 \text{Wohnfläche}_i + \hat{\beta}_2 \text{Schlafzimmeranzahl}_i + \hat{\beta}_3 \text{Baujahr}_i$$

Was ist nun Ramsey's Idee?

1. Schätze das Ausgangsmodell:

$$\text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i + \epsilon_i$$

2. Berechne die Prognosen:

$$\color{red}{\widehat{\text{Hauspreis}}_i} = \hat{\beta}_0 + \hat{\beta}_1 \text{Wohnfläche}_i + \hat{\beta}_2 \text{Schlafzimmeranzahl}_i + \hat{\beta}_3 \text{Baujahr}_i$$

3. Schätze erneut, aber diesmal **mit Potenzen der Prognosen als Regressoren**:

$$\begin{aligned}\text{Hauspreis}_i = {} & \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i \\ & + \color{blue}{\beta_4} \color{red}{\widehat{\text{Hauspreis}}_i}^2 + \color{blue}{\beta_5} \color{red}{\widehat{\text{Hauspreis}}_i}^3 + \epsilon_i\end{aligned}$$

Was ist nun Ramsey's Idee?

1. Schätze das Ausgangsmodell:

$$\text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i + \epsilon_i$$

2. Berechne die Prognosen:

$$\widehat{\text{Hauspreis}_i} = \hat{\beta}_0 + \hat{\beta}_1 \text{Wohnfläche}_i + \hat{\beta}_2 \text{Schlafzimmeranzahl}_i + \hat{\beta}_3 \text{Baujahr}_i$$

3. Schätze erneut, aber diesmal **mit Potenzen der Prognosen als Regressoren**:

$$\begin{aligned} \text{Hauspreis}_i = \beta_0 + \beta_1 \text{Wohnfläche}_i + \beta_2 \text{Schlafzimmeranzahl}_i + \beta_3 \text{Baujahr}_i \\ + \beta_4 \widehat{\text{Hauspreis}_i}^2 + \beta_5 \widehat{\text{Hauspreis}_i}^3 + \epsilon_i \end{aligned}$$

4. Teste $H_0 : \beta_4 = \beta_5 = 0$ mit einem F-Test.

Warum nicht auch die erste Potenz der Prognosen?

Die $\widehat{\text{Hauspreis}}$ Werte sind bereits eine Linearkombination der Regressoren. Würden sie zusätzlich aufgenommen werden, würde das zu Multikollinearität führen.

Warum nicht auch die erste Potenz der Prognosen?

Die $\widehat{\text{Hauspreis}}$ Werte sind bereits eine Linearkombination der Regressoren. Würden sie zusätzlich aufgenommen werden, würde das zu Multikollinearität führen.

Bis zur wievielten Potenz?

Dafür gibt es keine allgemeine Regel, meistens wählt man die 2. und 3. Potenz.

Warum nicht auch die erste Potenz der Prognosen?

Die $\widehat{\text{Hauspreis}}$ Werte sind bereits eine Linearkombination der Regressoren. Würden sie zusätzlich aufgenommen werden, würde das zu Multikollinearität führen.

Bis zur wievielten Potenz?

Dafür gibt es keine allgemeine Regel, meistens wählt man die 2. und 3. Potenz.

Vor- und Nachteile des Verfahrens?

- Kann auch bei einer Vielzahl an Regressoren effizient angewendet werden. Wir brauchen kein Alternativmodel definieren.
- Es kann zwar aufgezeigt werden, *ob* eine Fehlspezifikation vorliegt, aber nicht *welcher Art*.