AMS 7
More on Regression
Lecture 15

Department of Applied Mathematics and Statistics, University of California, Santa Cruz

Winter 2014
How good is a regression model?

- Statistical significance - test if $\beta_1 = 0$
- Practical significance - $r^2$
- Check model assumptions - residual plots
Hypothesis Testing for Regression

★ The model is $y = \beta_0 + \beta_1 x$, where $\beta_0$ and $\beta_1$ are population parameters.

→ If there is a linear relationship between $x$ and $y$ then $\beta_1 \neq 0$.

♣ This is a t-test with $n - 2$ degrees of freedom.

1. $H_0$: $\beta_1 = 0$ vs. $H_1$: $\beta_1 \neq 0$
2. Level of significance $\alpha = 0.05$
3. Test statistic: $t = \frac{b_1 - 0}{s_{b_1}}$ (sampling distribution under $H_0$ is $t$ with $n - 2$ df)
4. Compute $t$ and its p-value with JMP
5. Reject if p-value $< 0.05$
6. Draw conclusions about linear relationship
\[ r^2 = \text{square of correlation between } x \text{ and } y \]

= \% of variability in \( y \) is explained by predicting from \( x \)

\[ = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \]

= explained variation

\[ \text{total variation} \]

\[ \rightarrow \text{Recall that } s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n}(y_i - \bar{y})^2 \]

- \( 0 \leq r^2 \leq 1 \)

- Gives a measure for practical significance
Model assumptions:

1. $y$ is normally distributed with mean $\beta_0 + \beta_1 x$ and standard deviation $\sigma$.
2. The relationship between $x$ and $y$ is linear.
3. $\sigma$ is the same for all observations.
4. The observation $(x_i, y_i)$ is independent of $(x_j, y_j)$ (conditional on $\beta_0, \beta_1$)

How do we check these?

→ Hypothesis test for (1) and (2).
→ Residual analysis for (2), (3) and (4).
**Residuals**: \( e_i = y_i - \hat{y}_i \)

→ Plot \( x_i \) vs. \( e_i \) or \( \hat{y}_i \) vs. \( e_i \) (BUT not \( y_i \) vs. \( e_i \), which are correlated)

→ Make sure there are no patterns in the plot
  
  • check for non-linearity
  
  • check for change in variability (heteroscedasticity)

**Patterns indicate violations of assumptions!**

Prediction is *valid* only when statistically significant and no problems with residuals.

**Prediction interval**: a confidence interval for a predicted value - get from JMP.
Key Concepts!!!!!

- Test for linear relationship
- $r^2$
- Residual analysis