STAT 308 – Section 14.4/Chapter 15

```r
# This line of code tells the document all the display defaults

knitr::opts_chunk$set(echo = TRUE, cache = TRUE, include = TRUE, fig.align="center",warnings = FALSE,fi
```

Background Information

We have discussed multiple linear regression methods, where we fit a linear regression model to a particular response variable of interest to a variety of explanatory variables, both numeric and categorical. We have discussed methods of checking to see if the assumptions of homoscedasticity and normally distributed residuals are met. We will now discuss methods of how we can still perform linear regression analysis even when these assumptions are not met.

Response Variable Transformation

Motivating Example

Consider the dataset `Boston` available in the package `MASS`

With multiple linear regression, we have discussed the various assumptions needed to perform regression analysis. Combining these assumptions, with a response variable Y and p explanatory variables, $X_1, X_2, \ldots, X_p$, we have that

$$Y \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j, \sigma^2\right),$$

where we assume that all of the values are normally distributed with a linear mean structure and common variance.

However, let's perform a linear regression analysis for the median value of owner-occupied homes with crime rate as a covariate and check to see if the assumptions of homoscedasticity and normally distributed residuals are met.

Transformation: Inputting the response variable (Y) into a function to obtain a new response ($\tilde{Y}$) that more closely meets the assumptions for a linear regression model.

Common Types of transformations of a response variable:

- Log-Transformation ($\tilde{Y} = \log(Y), Y > 0$):
  - stabilizes the variance if the variability increases significantly with any X
  - normalizes the variable if it is highly right skewed (many points on the upper end of QQ-plot are above the 45-degree line)

  - linearizes the relationship if the relationship between X and Y appears to be exponential

- Square Root-Transformation ($\tilde{Y} = \sqrt{Y}, Y \geq 0$):
  - stabilizes the variance if the variance is proportional to the mean (i.e. $\text{sd}(Y|X_1, \dots, X_p) = c\text{E}(Y|X_1, \dots, X_p)$ where c is a constant)
- Squared-Transformation ($\tilde{Y} = Y^2$):
  - stabilizes the variance if the variance is decreases significantly with any X

The method of finding the regression estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ is still minimizing the sum of squared errors:

$$\sum_{i=1}^{n} \left(\tilde{y}_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_j \right)^2$$

Example

Perform a least squares regression analysis for the natural log of median home value with crime rate as the only explanatory variable by answering the following questions:

a. Write out the least squares regression line:

b. Interpret the slope of the least squares regression line in the context of the given problem.

c. Perform an F-test to determine a linear model with crime rate significantly improves the predictive ability of log sales price.

d. Check to see if the assumptions of homoscedasticity and normally distibuted residuals are met for this transformed linear model.

Polynomial Regression

The scatterplot for the log transformation appears to fit the data better, but there does seem to be a little bit of curvature of the mean curve (red line) in the residual plot. How can we adjust our model to account for this curvature?

Polynomial Regression: Extension of a linear regression model where we now allow for quadratic, cubic, and higher-order terms in our regression model. If we have a single explanatory variable X, then we can have

$$Y \sim \mathcal{N}\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p, \sigma^2\right)$$

or if Y is transformed

$$\tilde{Y} \sim \mathcal{N}\left(\beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p, \sigma^2\right)$$

The method of finding the regression estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ is still minimizing the sum of squared errors:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_i^j \right)^2$$

Example

Perform a least squares regression analysis for log median home value with a linear and quadratic term for crime rate by answering the following questions:

a. Write out the least squares regression line:

b. Perform an F-test to determine if a polynomial model with a linear and quadratic term significantly improves the predictive ability of log median home value.

c. Perform an F-test to determine if adding a quadratic term to a least squares regression line which already includes a linear term significantly improves the predictive ability of log median home value.

d. Check to see if the assumptions of homoscedasticity and normally distibuted residuals are met for this transformed polynomial model.

Example

Perform an F-test to determine if adding a cubic term to a least squares regression line which already includes linear and quadratic terms significantly improves the predictive ability of log median home value.