Lecture 12:
Diagnostics for Multiple Linear Regression
STAT 632, Spring 2020

Hat Matrix

The $n \times 1$ vector of fitted values is given by:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$

- The $n \times n$ matrix $\boldsymbol{H}$ is called the **hat matrix** since it transforms the vector of observed responses $\boldsymbol{Y}$ into the vector of fitted responses $\hat{\boldsymbol{Y}}$
- $\boldsymbol{H}$ is an *idempotent* matrix since $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$
- $\boldsymbol{H}$ is a symmetric matrix since $\boldsymbol{H} = \boldsymbol{H}'$

Properties of Residuals in MLR

For a multiple linear regression model, the vector of residuals is given by

$$\hat{\boldsymbol{e}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$$

The expected value of the vector of residuals:

$$E(\hat{\boldsymbol{e}}) = \boldsymbol{0}$$

The $n \times n$ variance-covariance matrix of the vector of residuals:

$$Var(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

Derivation provided in Sheather, Section 6.1.2, p. 154

Properties of Residuals in MLR

- For the MLR model we assume that $Var(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$, so the errors are uncorrelated an have constant variance.
- However, we found that $Var(\hat{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$, so the residuals do not have constant variance and are correlated.
- Fortunately, the impact of this is usually small, and regression diagnostics can be applied using the either the raw residuals, or by standardizing the residuals by their standard errors.

Assumptions for MLR

- **Linearity**: Y is related to predictors $x_1, \cdots, x_p$ by a multiple linear regression model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e$ with mean $E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p$. That is, Y can be modeled as a linear combination of the predictors.
- **Independence**: The errors $e_1, e_2, \cdots, e_n$ are independent of each other; $Cov(e_i, e_j) = 0$ when $i \neq j$.
- **Constant Variance**: The errors $e_1, e_2, \cdots, e_n$ have common variance $Var(e_i) = \sigma^2$.
- **Normality**: The errors follow a normal distribution; $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

Potential Problems

- Nonlinear relationships between the response and the predictors that are not accounted for by the model.
- Moderate to severe nonconstant variability in the residuals (heteroscedasticity).
- Outliers and high leverage points.
- Collinearity among the predictor variables.

We can use **regression diagnostics** to check the validity of the regression model and evaluate any potential problems.

Leverage Points

- The leverage for point i is quantified by h_i, the i^{th} diagonal entry of hat matrix $\boldsymbol{H}$.
- Intuitively, a high leverage point has extreme or unusual values for the predictors, when compared to the bulk of the data.
- A popular rule is to classify the i^{th} point as a point of high leverage in a multiple linear regression model with p predictors if

$$h_i > 2 \times \text{average}(h_i) = \frac{2(p+1)}{n}$$

- Note that $\sum_{i=1}^{n} h_i = p + 1$

Standardized Residuals

The variance of the i^{th} residual is given by

$$Var(\hat{e}_i) = \sigma^2(1 - h_i)$$

where h_i is the i^{th} diagonal entry of $\boldsymbol{H}$.

Thus, the i^{th} standardized residual, r_i, is given by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

where $\hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i}{n-p-1}}$ is the residual standard error.

Identifying Outliers

- Recall, an **outlier** is a point that has a response value (y_i) that does not follow the trend set by the bulk of the data.
- We can classify a point as an outlier if its standardized residual falls outside the interval from **-2 to 2**. For large data sets, change this rule to **-4 to 4** (otherwise, too many points would be flagged).
- Just because a point is an outlier and/or has high leverage does not mean we must ignore that point and remove it from the model. Rather, outliers and/or high leverage points should be investigated, and can provide important insights about the data. Sometimes outliers and/or leverage points indicate a problem with the data that can be corrected.

Residual Plots

- Residual plots are one of the most useful diagnostics for a multiple linear regression model.
- The most important diagnostic is a plot of the residuals, $\hat{e}_i$, versus the fitted values, $\hat{y}_i$. Alternatively, we can use the standardized residuals, r_i, which are useful for outlier detection.
- It is also worthwhile to make a plot of the residuals, $\hat{e}_i$, versus each predictor variable. Again, alternatively, we can use the standardized residuals, r_i, instead of the raw residuals.
- Ideally, the residual plots should show no obvious patterns or nonconstant variability, and the points are randomly scattered around 0.

Example: Menu Pricing Data Set

Recall, the data set from Zagat surveys of customers of 168 Italian restaurants in New York City. We considered the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

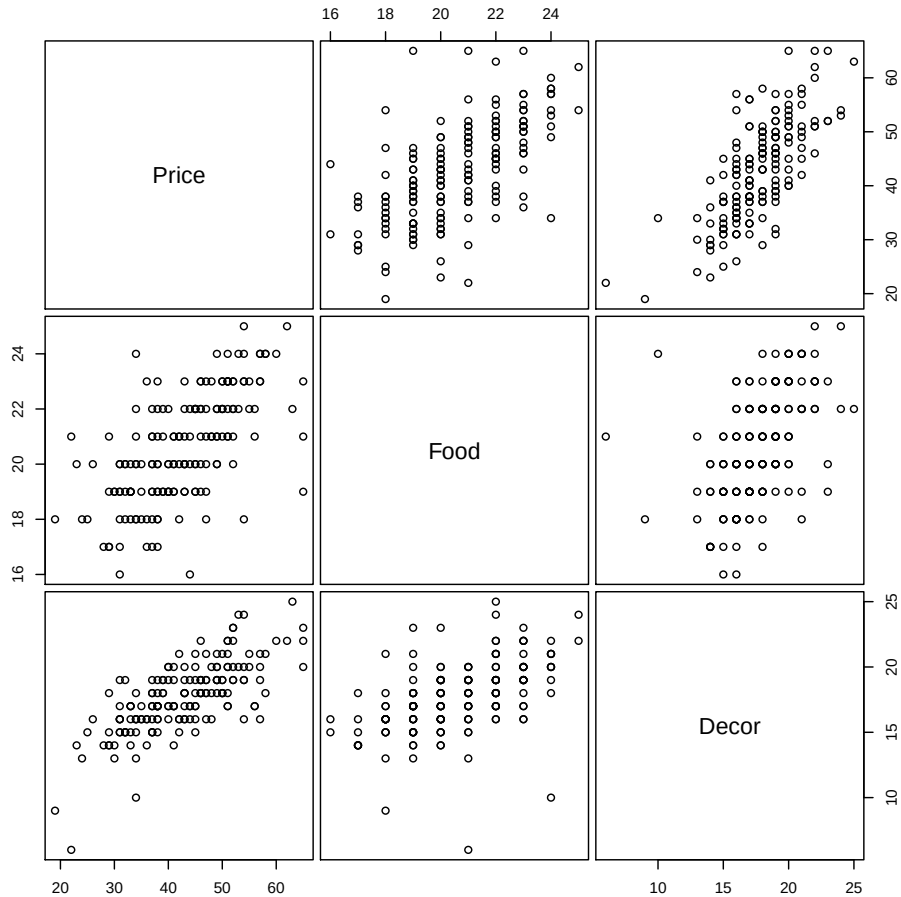
- Y = Price = the price (in \$US) of dinner (including 1 drink and tip)
- x_1 = Food = customer rating of the food (out of 30)
- x_2 = Decor = customer rating of the decor (out of 30)
- x_3 = East = dummy variable, 1 (0) if the restaurant is east (west) of Fifth Avenue

An additional predictor `Service` was removed since it was not significant.

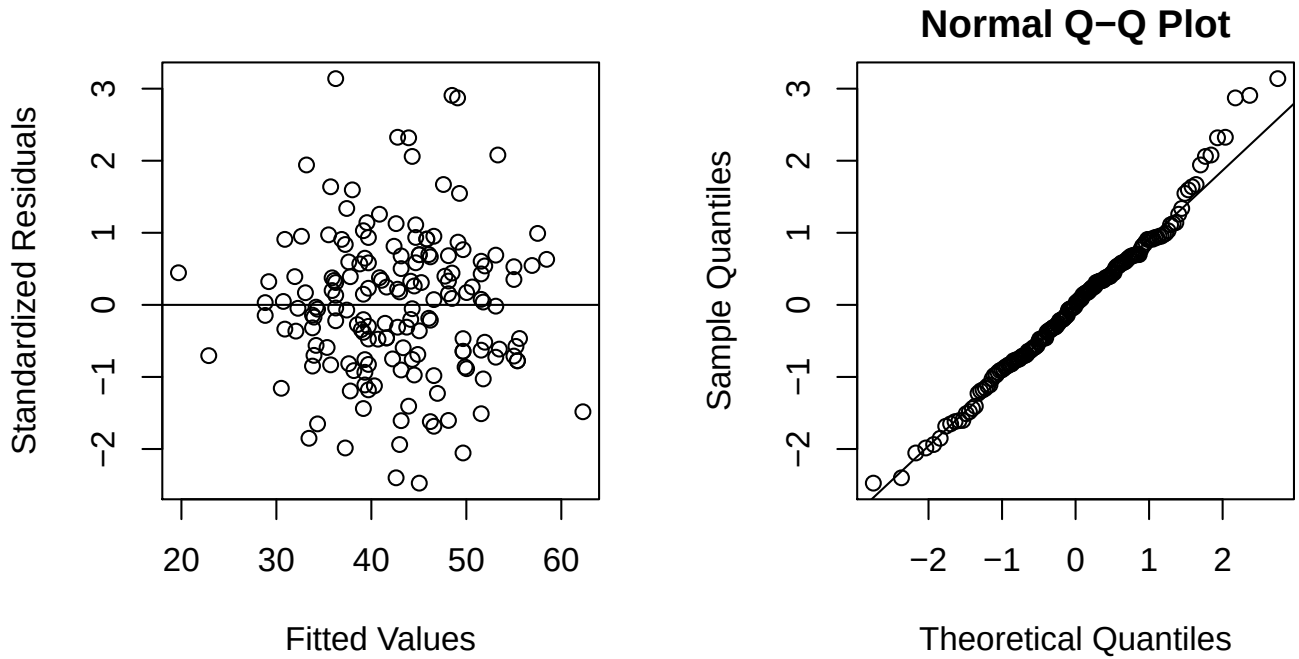
```
> nyc <- read.csv("https://ericwfox.github.io/data/nyc.csv")
> lm2 <- lm(Price ~ Food + Decor + East, data=nyc)
> summary(lm2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279,Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

The scatter plot matrix shows that the predictor variables have linear relationships with the response.

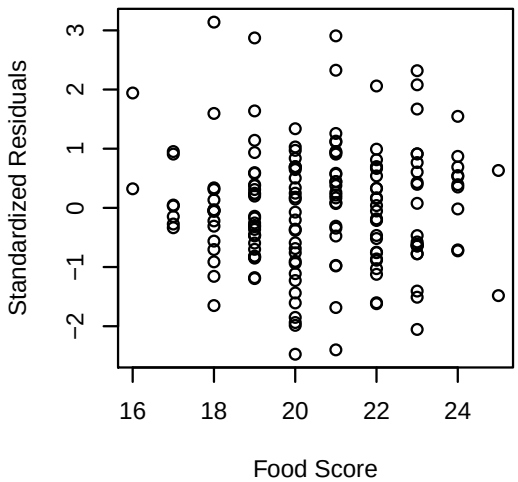
```
> pairs(Price ~ Food + Decor, data=nyc)
```

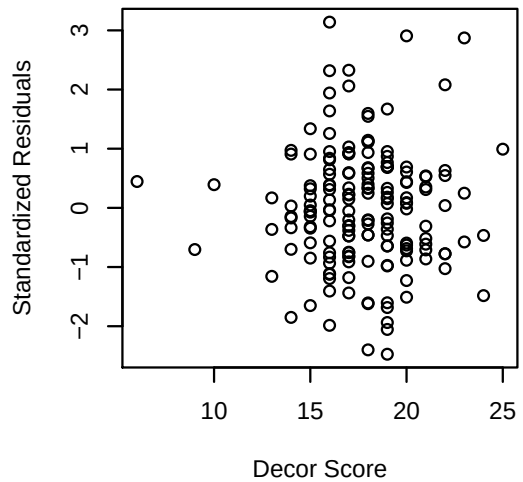


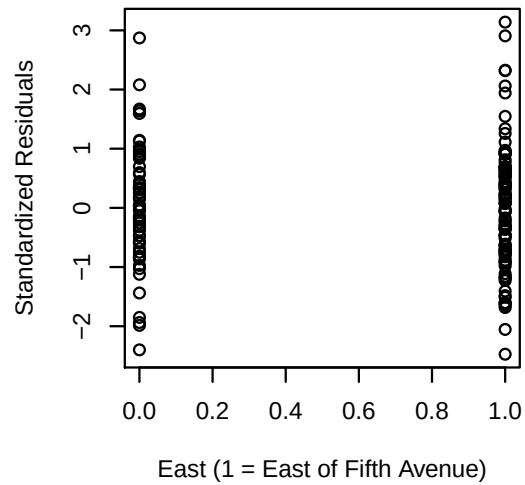
The plot of the standardized residuals versus fitted values shows no discernible trend or nonconstant variance – the points are randomly scattered around 0. The assumptions of linearity and nonconstant variance appear satisfied. The QQ plot also indicates that distribution of the standardized residuals are approximately normal, and that there are no extreme outliers.



The plots of the residuals versus each predictor also indicate that the MLR assumptions are satisfied.



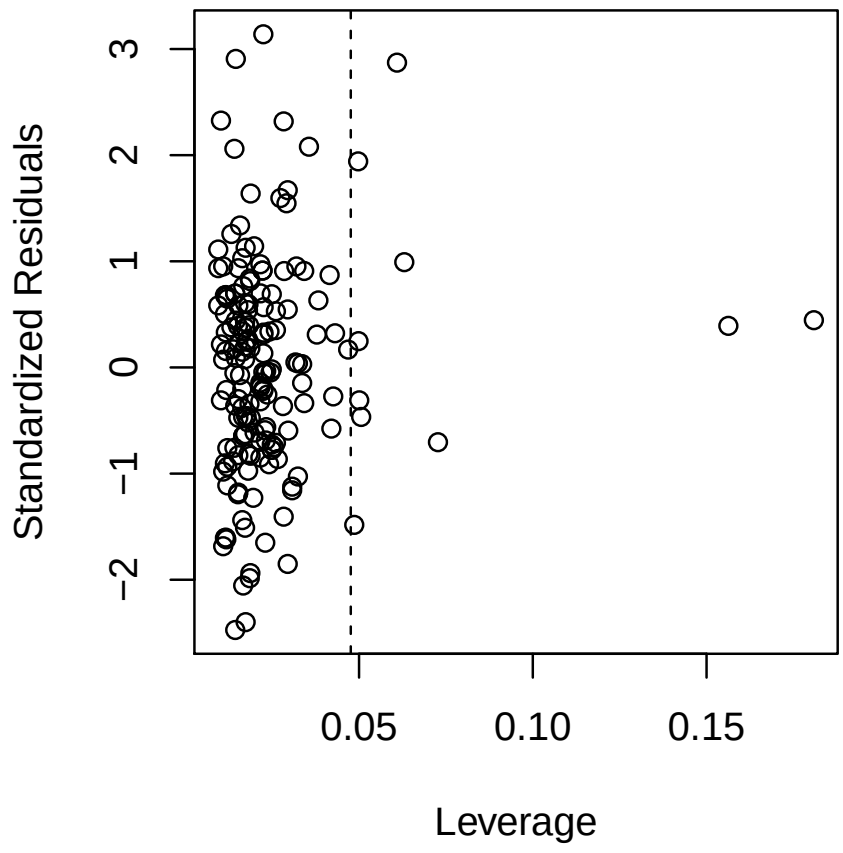




Here is the code for the diagnostic plots:

```
# residuals versus fitted and QQ plot
> par(mfrow=c(1,2), mar=c(4.5, 4.5, 2, 2))
> plot(predict(lm2), rstandard(lm2),
       xlab="Fitted Values", ylab="Standardized Residuals")
> abline(h=0)
> qqnorm(rstandard(lm2))
> qqline(rstandard(lm2))

# residuals versus predictors
> par(mfrow=c(1,3), mar=c(4.5, 4.5, 2, 2))
> plot(nyc$Food, rstandard(lm2),
       xlab="Food Score", ylab="Standardized Residuals")
> plot(nyc$Decor, rstandard(lm2),
       xlab="Decor Score", ylab="Standardized Residuals")
> plot(nyc$East, rstandard(lm2),
       xlab="East (1 = East of Fifth Avenue)",
       ylab = "Standardized Residuals")
```

```
> p <- 3
> n <- nrow(nyc)
> plot(hatvalues(lm2), rstandard(lm2),
       xlab='Leverage', ylab='Standardized Residuals')
> abline(v = 2*(p+1)/n, lty=2)
```



Your Turn

Idenitfy the two restaurants with the highest leverages.