Lecture 15:
Categorical Predictors and Interactions
STAT 632, Spring 2024

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 $\sum_{i=1}^{n} h_i = p + 1$ $\rightarrow$ sum of the main diagonal for a projection matrix is equal to the rank

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 $\hat{\sigma} = \sqrt{\frac{RSS}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-p-1}}$ 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:

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

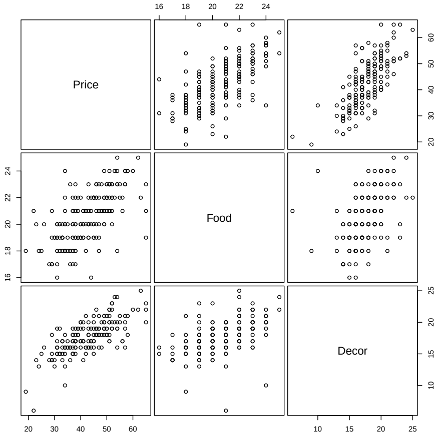
- Y = Price = the price (in $US) of dinner (including 1 drink and tip)
- x_1 = Food = customer rating of the food (out of 30)
- x_2 = Decor = customer rating of the decor (out of 30)
- x_3 = East = dummy variable, 1 (0) if the restaurant is east (west) of Fifth Avenue

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

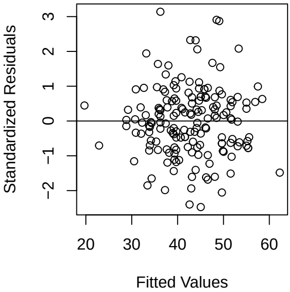
```
> nyc <- read.csv("nyc.csv")
> lm2 <- lm(Price ~ Food + Decor + East, data=nyc)
> summary(lm2)

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
```

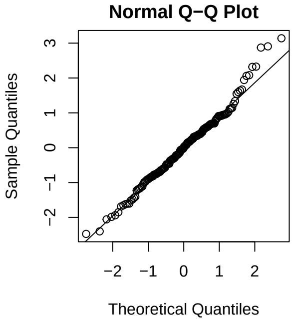
The scatter plot matrix shows that the predictor variables have linear relationships with the response.

```
> 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 constant variance appear satisfied. The QQ plot also indicates that distribution of the standardized residuals are approximately normal, and that there are no extreme outliers.



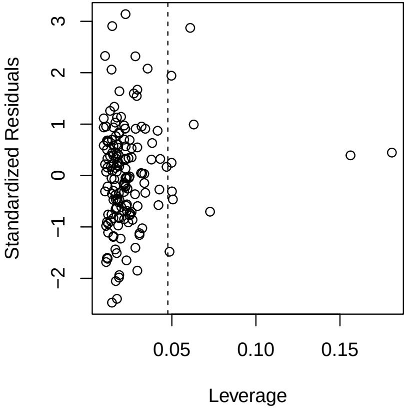


The plots of the residuals versus each predictor also indicate that the MLR assumptions are satisfied.

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.

Introduction

- Predictors in a multiple linear regression model can either be *quantitative* (e.g, weight, age) or *qualitative* (e.g., gender, education level). Qualitative predictors are also called *categorical* or *factors*.
- A categorical predictor with two levels (0 or 1) is called a *dummy* or *indicator* variable. as.factor( )
- Sometimes the effect that a quantitative predictor has on the response changes depending on the level of categorical predictor. For example, perhaps the effect age has on salary depends on the education status of the person. This is called an *interaction* effect.

Parallel Regression Lines

Let x be a quantitative variable, and d a dummy variable.

$$Y = \beta_0 + \beta_1 x + \beta_2 d + \epsilon = \begin{cases} \beta_0 + \beta_1 x + \epsilon, & \text{if } d = 0 \\ (\beta_0 + \beta_2) + \beta_1 x + \epsilon, & \text{if } d = 1 \end{cases}$$

- This model gives two separate regression lines that have the same slope but different intercepts.
- The parameter β_2 represents the vertical distance between the two lines.

Unrelated Regression Lines

Let x be a quantitative variable, and d a dummy variable.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x + \beta_2 d + \beta_3 \overbrace{d \cdot x}^{\text{interaction}} + \epsilon \\
&= \begin{cases} \beta_0 + \beta_1 x + \epsilon, & \text{if } d = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x + \epsilon, & \text{if } d = 1 \end{cases}
\end{aligned}
$$

- This model gives two separate regression lines that have different slopes, and different intercepts.
- β_3 is the coefficient for the *interaction* between the dummy variable, d, and the quantitative variable, x.

Example: Credit Card Data Set

- We consider the Credit data set from the ISLR package. Type help(Credit) to read about this data set in the help menu.
- The response variable is Balance, the average credit card balance in dollars.
- The predictors of interest are Income (in thousands of dollars) and Student, a dummy variable indicating student status (No $= 0$ or Yes $= 1$).

```
> library(ISLR)
> head(Credit, n=5)
  ID  Income Limit Rating Cards Age Education Gender Student Married Ethnicity Balance
1  1  14.891  3606    283     2  34        11   Male      No     Yes Caucasian     333
2  2 106.025  6645    483     3  82        15 Female     Yes     Yes     Asian     903
3  3 104.593  7075    514     4  71        11   Male      No      No     Asian     580
4  4 148.924  9504    681     3  36        11 Female      No      No     Asian     964
5  5  55.882  4897    357     2  68        16   Male      No     Yes Caucasian     331
```

```
> lm1 <- lm(Balance ~ Income + Student, data=Credit)

# shows coding R uses for the dummy variable
> contrasts(Credit$Student)
    Yes
No    0
Yes   1

> summary(lm1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.1430    32.4572   6.505 2.34e-10 ***
Income        5.9843     0.5566  10.751  < 2e-16 ***
StudentYes  382.6705    65.3108   5.859 9.78e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 391.8 on 397 degrees of freedom
Multiple R-squared:  0.2775, Adjusted R-squared:  0.2738
F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

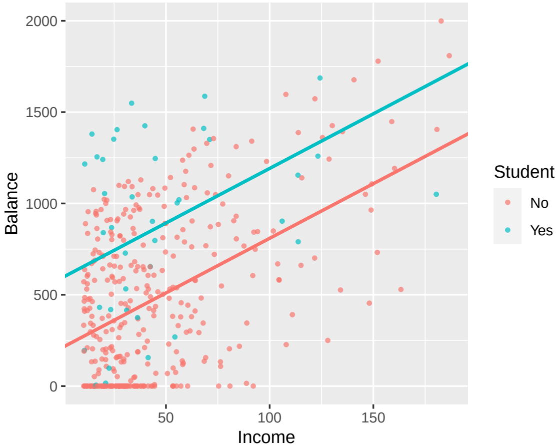
We can write the regression equation for the fit:

$$\begin{aligned}\widehat{\texttt{Balance}} &= \hat{\beta}_0 + \hat{\beta}_1\texttt{income} + \hat{\beta}_2\texttt{student} \\ &= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1\texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1\texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}\end{aligned}$$

Plugging in the coefficients from the regression summary gives:

$$\begin{aligned}\widehat{\texttt{Balance}} &= 211.14 + 5.98\texttt{income} + 382.67\texttt{student} \\ &= \begin{cases} 211.14 + 5.98\texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ 593.81 + 5.98\texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}\end{aligned}$$

```
ggplot(Credit, aes(Income, Balance, colour = Student)) +
  geom_point(alpha=0.7) +
  geom_abline(intercept = 211.1, slope = 5.98, colour = "#F8766D") +
  geom_abline(intercept = 593.8, slope = 5.98, colour = "#00BFC4")
```



```
> lm2 <- lm(Balance ~ Income + Student + Income:Student, data=Credit)
> summary(lm2)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       200.6232    33.6984   5.953 5.79e-09 ***
Income              6.2182     0.5921  10.502  < 2e-16 ***
StudentYes        476.6758   104.3512   4.568 6.59e-06 ***
Income:StudentYes  -1.9992     1.7313  -1.155    0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799, Adjusted R-squared:  0.2744
F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

We can write the regression equation for the fit:

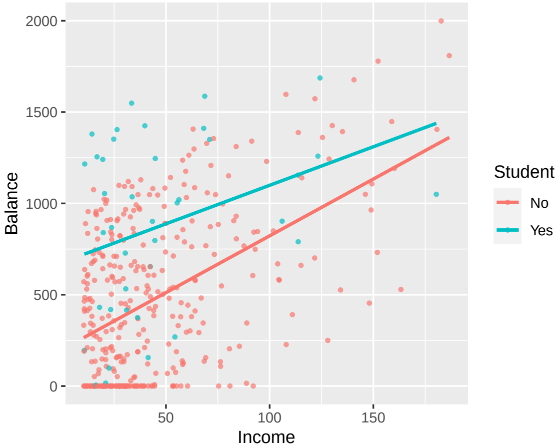
$$\widehat{\texttt{Balance}} = \hat{\beta}_0 + \hat{\beta}_1 \texttt{income} + \hat{\beta}_2 \texttt{student} + \hat{\beta}_3 \texttt{student} \cdot \texttt{income}$$
$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}$$

Plugging in the coefficients from the regression summary gives:

$$\widehat{\texttt{Balance}} = 200.62 + 6.22 \texttt{income} + 476.68 \texttt{student} - 2.00 \texttt{student} \cdot \texttt{income}$$
$$= \begin{cases} 200.62 + 6.22 \texttt{income}, & \text{if } \texttt{student}{=}0 \text{ (No)} \\ 677.3 + 4.22 \texttt{income}, & \text{if } \texttt{student}{=}1 \text{ (Yes)} \end{cases}$$

Note that the coefficient for the interaction, β_3, is not significant (p-value$= 0.249$), so we do not necessarily need to include the interaction term.

```
ggplot(Credit, aes(Income, Balance, colour = Student)) +
  geom_point(alpha=0.7) +
  geom_smooth(method="lm", se=FALSE)
```



Categorical Predictors with More Than Two Levels

- When a categorical predictor contains more than two levels, we create additional dummy variables. costs more df
- For example, consider the `Wage` data set also from the `ISLR` package. The data contain information on 3000 males workers in the Mid-Atlantic region.
- The response variable is `logwage`, the log of the workers wage.
- The predictor `education` is a categorical variable indicating education level with 5 levels: 1. $<$ HS Grad, 2. HS Grad, 3. Some College, 4. College Grad, and 5. Advanced Degree.

$\nearrow$ 5-1

We can write the regression equation with 4 dummy variables:

$$
\begin{aligned}
\log(\texttt{Wage}) &= \beta_0 + \beta_1 \texttt{HS\_Grad} + \beta_2 \texttt{Some\_College} \\
&\quad + \beta_3 \texttt{College\_Grad} + \beta_4 \texttt{Advanced\_Degree} + \epsilon \\
&= \begin{cases}
\beta_0 + \epsilon & \text{if } {<}\texttt{HS\_Grad} \text{ (baseline)} \\
\beta_0 + \beta_1 + \epsilon & \text{if } \texttt{HS\_Grad} = 1 \\
\beta_0 + \beta_2 + \epsilon & \text{if } \texttt{Some\_College} = 1 \\
\beta_0 + \beta_3 + \epsilon & \text{if } \texttt{College\_Grad} = 1 \\
\beta_0 + \beta_4 + \epsilon & \text{if } \texttt{Advanced\_Degree} = 1
\end{cases}
\end{aligned}
$$

In general, if we have a categorical variable with k levels, then the regression equation contains $k - 1$ dummy variables.

β_1 represents the "increase" in logwage for someone who graduated highschool versus someone who didn't.

```
ggplot(Wage, aes(education, logwage)) +
  geom_boxplot() + coord_flip()
```