



THE UNIVERSITY OF TEXAS AT AUSTIN  
McCOMBS SCHOOL OF BUSINESS

# Diagnostics & Transformations 2

---

## Lecture 18

STA 371G

## Salaries of newly hired managers



## Salaries of newly hired managers



- Salary (response)
- Manager rating
- Years of experience
- Years since graduation
- Origin (internal or external hire)

## Data issues

Data scientists report that they spend 70% of their time on obtaining and cleaning the data. Only 30% is for statistical analysis.

## Data issues

Data scientists report that they spend 70% of their time on obtaining and cleaning the data. Only 30% is for statistical analysis.

Never run a regression without exploring and cleaning the data first!

The most common issues:

1. Outliers
2. Missing data
3. Multicollinearity
4. Assumption violations
5. High-leverage points

1. Outliers

2. Missing data

3. Multicollinearity

4. Assumption violations

5. High-leverage points

## Exploring the data: Outliers

Boxplots are commonly used to detect outliers. Let's start by looking at the Salary column.



## Exploring the data: Outliers

Boxplots are commonly used to detect outliers. Let's start by looking at the Salary column.

```
boxplot(manager$Salary, xlab='Salary', horizontal=T)
```



## Exploring the data: Outliers

```
manager[manager$Salary>200,]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
146	511	6.1	2	2	Internal

```
manager[manager$Salary<0,]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
121	-66	5.7	1	2	Internal



## Exploring the data: Outliers

```
manager[manager$Salary>200,]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
146	511	6.1	2	2	Internal

```
manager[manager$Salary<0,]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
121	-66	5.7	1	2	Internal

We can deal with outliers in two ways.

- If the result of **errors in the data**, we can try to correct or omit.



## Exploring the data: Outliers

```
manager[manager$Salary>200,]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
146	511	6.1	2	2	Internal

```
manager[manager$Salary<0,]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
121	-66	5.7	1	2	Internal

We can deal with outliers in two ways.

- If the result of **errors in the data**, we can try to correct or omit.
- If not, consider omitting, but report on them separately.



## Exploring the data: Outliers

Let's omit the outliers by creating a new data set `mclean` that consists of the subset of the data where the salary is between \$0 and \$200,000.

```
mclean <- manager[manager$Salary>0 &  
                  manager$Salary<200,]
```



## Exploring the data: Outliers

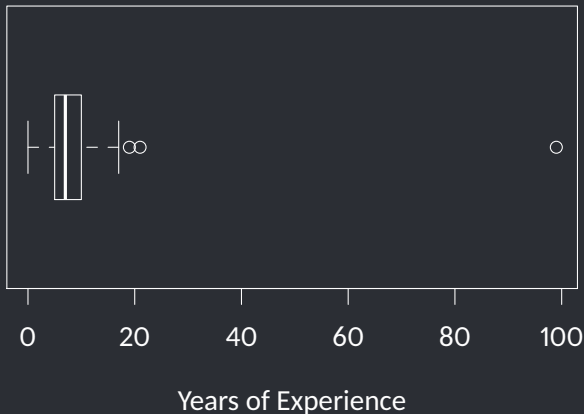
Let's omit the outliers by creating a new data set `mclean` that consists of the subset of the data where the salary is between \$0 and \$200,000.

```
mclean <- manager[manager$Salary>0 &  
                  manager$Salary<200,]
```



## Exploring the data: Outliers

```
boxplot(mclean$YearsExp, xlab='Years of Experience',  
        horizontal=T)
```



## Exploring the data: Outliers

99 must be a code for missing entry in the Years of Experience variable!



## Exploring the data: Outliers

99 must be a code for missing entry in the Years of Experience variable!

Let's label all 99s as NA (Not Available — R's code for missing data).

## Exploring the data: Outliers

99 must be a code for missing entry in the Years of Experience variable!

Let's label all 99s as NA (Not Available — R's code for missing data).

```
mclean$YearsExp[mclean$YearsExp == 99] <- NA
```

1. Outliers

2. Missing data

3. Multicollinearity

4. Assumption violations

5. High-leverage points

## Exploring the data: Missing entries

Let's see if we have other missing data.

## Exploring the data: Missing entries

Let's see if we have other missing data.

```
mclean[!complete.cases(mclean),]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
103	75	NA	8	8	Internal
110	81	NA	9	9	External
124	73	5.9	NA	7	External
154	49	8.0	1	1	<NA>

## Exploring the data: Missing entries

Let's see if we have other missing data.

```
mclean[!complete.cases(mclean),]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
103	75	NA	8	8	Internal
110	81	NA	9	9	External
124	73	5.9	NA	7	External
154	49	8.0	1	1	<NA>

This isn't surprising—it is very common to have missing entries in your data.

## Exploring the data: Missing entries

There are two ways of dealing with missing data:

- Omit the rows that have missing entries in it.

Omitting data is the easiest, but often **not the best way, because you lose all the other information available in the same row.**

## Exploring the data: Missing entries

There are two ways of dealing with missing data:

- Omit the rows that have missing entries in it.
- Try to predict values to fill the missing entries.

Omitting data is the easiest, but often **not the best way, because you lose all the other information available in the same row.**



## Exploring the data: Missing entries

There are two ways of dealing with missing data:

- Omit the rows that have missing entries in it.
- Try to predict values to fill the missing entries.

Omitting data is the easiest, but often **not the best way, because you lose all the other information available in the same row.**

## Exploring the data: Missing entries

There are two ways of dealing with missing data:

- Omit the rows that have missing entries in it.
- Try to predict values to fill the missing entries.

Omitting data is the easiest, but often **not the best way, because you lose all the other information available in the same row.**

Let's try to fill in some estimates.

## Exploring the data: Missing entries

What should we replace the NAs in the Manager Rating and Years of Experience columns with?

## Exploring the data: Missing entries

What should we replace the NAs in the Manager Rating and Years of Experience columns with?

The simplest way would be to use the averages in the respective columns.

```
mclean$MngrRating[is.na(mclean$MngrRating)] <-  
  mean(mclean$MngrRating, na.rm=T)
```

```
mclean$YearsExp[is.na(mclean$YearsExp)] <-  
  mean(mclean$YearsExp, na.rm=T)
```

## Exploring the data: Missing entries

What should we replace the NAs in the Manager Rating and Years of Experience columns with?

The simplest way would be to use the averages in the respective columns.

```
mclean$MngrRating[is.na(mclean$MngrRating)] <-  
  mean(mclean$MngrRating, na.rm=T)  
  
mclean$YearsExp[is.na(mclean$YearsExp)] <-  
  mean(mclean$YearsExp, na.rm=T)
```

A smarter and more advanced way is to predict the missing data from the other data (using regression!).

## Exploring the data: Missing entries

What about the missing data for categorical variables?

## Exploring the data: Missing entries

What about the missing data for categorical variables? Let's choose the easy way and omit them.

```
mclean <- na.omit(mclean)
```

## Exploring the data: Missing entries

What about the missing data for categorical variables? Let's choose the easy way and omit them.

```
mclean <- na.omit(mclean)
```

This removes all the rows that contain missing entries (only the Origin column has missing entries in this case.)



## Exploring the data: Missing entries

What about the missing data for categorical variables? Let's choose the easy way and omit them.

```
mclean <- na.omit(mclean)
```

This removes all the rows that contain missing entries (only the Origin column has missing entries in this case.)

We could also predict the missing entries, or treat the missing entries as a separate level (e.g. "Unknown").

## Exploring the data: Missing entries

Important things to consider:

- While dealing with the missing data, we assume that the data is “Missing Completely at Random” (MCAR).

## Exploring the data: Missing entries

### Important things to consider:

- While dealing with the missing data, we assume that the data is “Missing Completely at Random” (MCAR).
- If this assumption does not hold (e.g. if the missing data mostly belongs to external hires), the model will be biased.

## Exploring the data: Missing entries

### Important things to consider:

- While dealing with the missing data, we assume that the data is “Missing Completely at Random” (MCAR).
- If this assumption does not hold (e.g. if the missing data mostly belongs to external hires), the model will be biased.
- Making predictions for missing data based on available data reinforces the existing relationships between variables, so impacts the standard error.

## Exploring the data: Missing entries

### Important things to consider:

- While dealing with the missing data, we assume that the data is “Missing Completely at Random” (MCAR).
- If this assumption does not hold (e.g. if the missing data mostly belongs to external hires), the model will be biased.
- Making predictions for missing data based on available data reinforces the existing relationships between variables, so impacts the standard error.
- If a lot of data is missing (e.g. more than 5%) for a particular variable, you may have to discard the whole column.

1. Outliers

2. Missing data

3. Multicollinearity

4. Assumption violations

5. High-leverage points

## Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

## Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors  $X_1$  and  $X_2$  are highly correlated, it is hard to estimate the effect of changing  $X_1$  while keeping  $X_2$  constant.



## Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors  $X_1$  and  $X_2$  are highly correlated, it is hard to estimate the effect of changing  $X_1$  while keeping  $X_2$  constant.
- This means we will have large standard errors, and large p-values, for  $X_1$  and  $X_2$ .

## Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors  $X_1$  and  $X_2$  are highly correlated, it is hard to estimate the effect of changing  $X_1$  while keeping  $X_2$  constant.
- This means we will have large standard errors, and large p-values, for  $X_1$  and  $X_2$ .
- This DOES NOT mean there isn't a relationship between  $X_1$  and  $Y$ , or  $X_2$  and  $Y$  – it just means we can't pin down that relationship, because of the correlation!

## Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors  $X_1$  and  $X_2$  are highly correlated, it is hard to estimate the effect of changing  $X_1$  while keeping  $X_2$  constant.
- This means we will have large standard errors, and large p-values, for  $X_1$  and  $X_2$ .
- This DOES NOT mean there isn't a relationship between  $X_1$  and  $Y$ , or  $X_2$  and  $Y$  – it just means we can't pin down that relationship, because of the correlation!

## Exploring the data: Multicollinearity

Multicollinearity exists whenever 2+ predictors in a regression model are moderately or highly correlated.

- If two predictors  $X_1$  and  $X_2$  are highly correlated, it is hard to estimate the effect of changing  $X_1$  while keeping  $X_2$  constant.
- This means we will have large standard errors, and large p-values, for  $X_1$  and  $X_2$ .
- This DOES NOT mean there isn't a relationship between  $X_1$  and  $Y$ , or  $X_2$  and  $Y$  – it just means we can't pin down that relationship, because of the correlation!

Correlation between the response and the predictors is good, but correlation between the predictors is not!

## Exploring the data: Multicollinearity

We want to avoid multicollinearity in our models!

## Exploring the data: Multicollinearity

We want to avoid multicollinearity in our models!

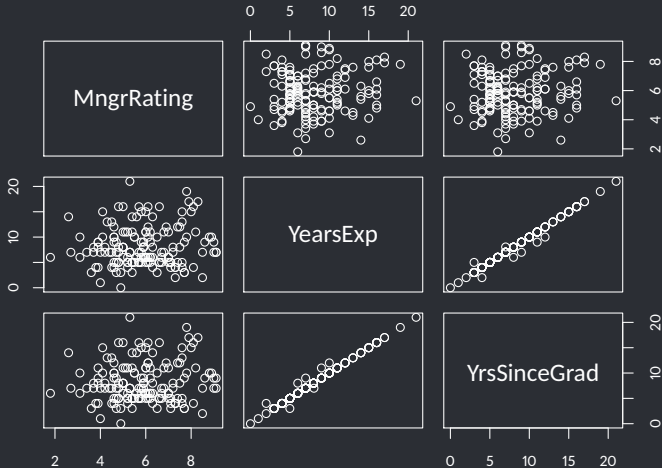
- Any conclusions based on the p-values, coefficients, and confidence intervals of the highly correlated variables will be unreliable.

## Exploring the data: Multicollinearity

We want to avoid multicollinearity in our models!

- Any conclusions based on the p-values, coefficients, and confidence intervals of the highly correlated variables will be unreliable.
- These statistics will not be stable: adding new data or predictors to the model could drastically change them.

```
pairs(~ MngrRating + YearsExp + YrsSinceGrad, data=mclean)
```





```
model <- lm(Salary ~ MngrRating + YearsExp + YrsSinceGrad + Origin,  
            data=mclean)  
summary(model)
```

Call:

```
lm(formula = Salary ~ MngrRating + YearsExp + YrsSinceGrad +  
    Origin, data = mclean)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-19.7766	-4.2842	-0.2906	3.3266	28.2773

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	54.1521	2.6071	20.771	< 2e-16	***
MngrRating	4.5147	0.3997	11.296	< 2e-16	***
YearsExp	-1.5262	1.3790	-1.107	0.270203	
YrsSinceGrad	0.7692	1.3833	0.556	0.578976	
OriginInternal	-4.7314	1.3878	-3.409	0.000838	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.838 on 149 degrees of freedom

Multiple R-squared: 0.6065, Adjusted R-squared: 0.596

F-statistic: 57.42 on 4 and 149 DF, p-value: < 2.2e-16



## Exploring the data: Multicollinearity

One way to see if two variables are collinear is to check the correlation between the two:

```
cor(mclean$YearsExp, mclean$YrsSinceGrad)
```

```
[1] 0.9947616
```

## Exploring the data: Multicollinearity

One way to see if two variables are collinear is to check the correlation between the two:

```
cor(mclean$YearsExp, mclean$YrsSinceGrad)  
  
[1] 0.9947616
```

Any correlation  $\geq 0.95$  is definitely a problem, but smaller correlations could be problematic too.

## Exploring the data: Multicollinearity

A better way to check multicollinearity is using Variance Inflation Factors (VIF).

- Remember: If we have multicollinearity, we will have large standard errors (i.e. high variance) in our estimators.
- The VIF is

$$\text{VIF}(\beta_j) = \frac{\text{Variance of } \beta_j \text{ in multiple regression}}{\text{Variance of } \beta_j \text{ if } X_j \text{ is the only predictor}}$$

- It is a measure of how much more uncertain we are about  $\beta_j$ , once we've included the other predictors.

## Exploring the data: Multicollinearity

```
library(car)
```

```
vif(model)
```

MngrRating	YearsExp	YrsSinceGrad	Origin
1.136002	95.954255	97.011260	1.540448

Predictors with VIF > 5 indicate multicollinearity.

## Exploring the data: Multicollinearity

```
library(car)
vif(model)
```

MngrRating	YearsExp	YrsSinceGrad	Origin
1.136002	95.954255	97.011260	1.540448

Predictors with  $VIF > 5$  indicate multicollinearity.

**Remember:** Multicollinearity could exist between more than two predictors (this is why there are only  $n - 1$  dummy variables for a categorical variable with  $n$  values).

## Exploring the data: Multicollinearity

A better way to check multicollinearity is using Variance Inflation Factors (VIF):

```
library(car)  
vif(model)
```

MngrRating	YearsExp	YrsSinceGrad	Origin
1.136002	95.954255	97.011260	1.540448

Drop any predictor that has  $VIF > 5$ .

## Exploring the data: Multicollinearity

A better way to check multicollinearity is using Variance Inflation Factors (VIF):

```
library(car)
vif(model)
```

MngrRating	YearsExp	YrsSinceGrad	Origin
1.136002	95.954255	97.011260	1.540448

Drop any predictor that has  $VIF > 5$ .

**Remember:** Multicollinearity could exist between more than two predictors (this is why there are only  $n - 1$  dummy variables for a categorical variable with  $n$  values).



```
model2 <- lm(Salary ~ MngrRating + YearsExp + Origin, data=mclean)
summary(model2)
```

Call:

```
lm(formula = Salary ~ MngrRating + YearsExp + Origin, data = mclean)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.8115	-4.3474	-0.3964	3.3358	28.1801

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	54.1080	2.5999	20.812	< 2e-16	***
MngrRating	4.5309	0.3977	11.394	< 2e-16	***
YearsExp	-0.7651	0.1687	-4.534	1.18e-05	***
OriginInternal	-4.6467	1.3762	-3.376	0.000935	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.823 on 150 degrees of freedom

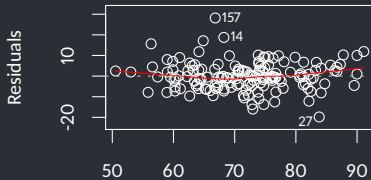
Multiple R-squared: 0.6057, Adjusted R-squared: 0.5978

F-statistic: 76.82 on 3 and 150 DF, p-value: < 2.2e-16

1. Outliers
2. Missing data
3. Multicollinearity
4. Assumption violations
5. High-leverage points

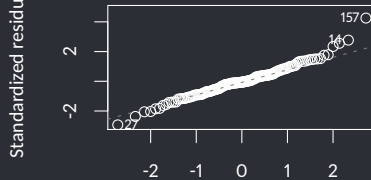
```
plot(model2)
```

Residuals vs Fitted



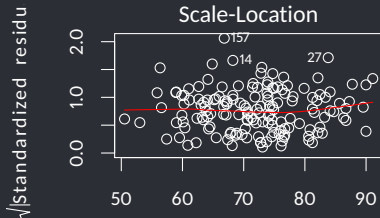
Fitted values

Normal Q-Q



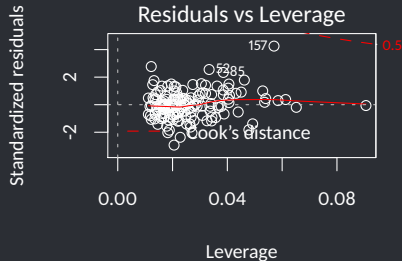
Theoretical Quantiles

Scale-Location



Fitted values

Residuals vs Leverage



1. Outliers
2. Missing data
3. Multicollinearity
4. Assumption violations
5. High-leverage points

## Outliers among the residuals

The data with index 157 seems to be hanging out by itself.

## Outliers among the residuals

The data with index 157 seems to be hanging out by itself.

We can display the indices of all of the outliers among the residuals.

```
boxplot(resid(model2))$out
```

14	27	52	85	157
18.76901	-19.81152	17.14133	15.66079	28.18007

## Outliers among the residuals

The data with index 157 seems to be hanging out by itself.

We can display the indices of all of the outliers among the residuals.

```
boxplot(resid(model2))$out
```

14	27	52	85	157
18.76901	-19.81152	17.14133	15.66079	28.18007

These cases are not predicted well by the model.

## Outliers among the residuals

Let's look at row 157:

```
manager[157, ]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
157	95	4	1	1	Internal

Someone with only 1 year of experience and a poor rating is hired as manager at \$95K!



## Outliers among the residuals

Let's look at row 157:

```
manager[157, ]
```

	Salary	MngrRating	YearsExp	YrsSinceGrad	Origin
157	95	4	1	1	Internal

Someone with only 1 year of experience and a poor rating is hired as manager at \$95K!

If you decide that this is an anomaly (e.g. the CEO's son was promoted!) that you don't want to include in your analysis, omit that row and report on it separately in your conclusions.

## Influential cases

- The Residuals vs Leverage plot tells about influential cases.

## Influential cases

- The Residuals vs Leverage plot tells about **influential cases**.
- A **high-leverage case** is one that has an unusual combination of predictor values.

## Influential cases

- The Residuals vs Leverage plot tells about **influential cases**.
- A **high-leverage case** is one that has an unusual combination of predictor values.
- An **influential case** is a high-leverage case that also has a high residual: it could change your  $\beta$  values significantly when excluded from your analysis, i.e., it does not follow the overall trend.

## Influential cases

- The Residuals vs Leverage plot tells about **influential cases**.
- A **high-leverage case** is one that has an unusual combination of predictor values.
- An **influential case** is a high-leverage case that also has a high residual: it could change your  $\beta$  values significantly when excluded from your analysis, i.e., it does not follow the overall trend.
- Look for the cases on the upper/lower right corners (beyond the dashed curves).