MSBA Boot Camp
Statistics Part II: Linear Regression

Weichen Wang

Innovation and Information Management
HKU Business School

[2] Some of the slides, figures, codes are from OpenIntro, Prof. Haipeng Shen, Dr. Mine Cetinkaya-Rundel, Dr. Wei Zhang and Dr. Dan Yang.

Roadmap

Boot Camp Stat I

- Single numerical variable
- Single categorical variable

Boot Camp Stat II

- Relationship between *multiple* numerical or categorical variables

Roadmap

Boot Camp Stat I

- Single numerical variable
- Single categorical variable

Boot Camp Stat II

- Relationship between *multiple* numerical or categorical variables
- *Predictive Analytics* is about using the data that is available to your company to predict potential future profitable opportunities.

Roadmap

Boot Camp Stat I

- Single numerical variable
- Single categorical variable

Boot Camp Stat II

- Relationship between *multiple* numerical or categorical variables
- *Predictive Analytics* is about using the data that is available to your company to predict potential future profitable opportunities.
  - Telecommunication companies predicting customer LTV
  - Customers understanding pricing of house/apartment
  - Insurance companies identifying fraudulent claims
  - Credit card companies predicting default possibility
  - Retail stores (Target) identifying pregnant customers

Roadmap

Boot Camp Stat I

- Single numerical variable
- Single categorical variable

Boot Camp Stat II

- Relationship between *multiple* numerical or categorical variables
- *Predictive Analytics* is about using the data that is available to your company to predict potential future profitable opportunities.
  - Telecommunication companies predicting customer LTV
  - Customers understanding pricing of house/apartment
  - Insurance companies identifying fraudulent claims
  - Credit card companies predicting default possibility
  - Retail stores (Target) identifying pregnant customers

$\Rightarrow$ *Regression and Classification*

Roadmap

Boot Camp Stat I

- Single numerical variable
- Single categorical variable

Boot Camp Stat II

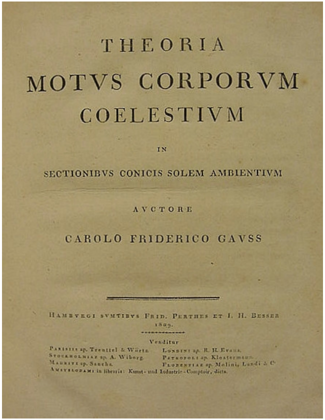
- Relationship between *multiple* numerical or categorical variables
- *Predictive Analytics* is about using the data that is available to your company to predict potential future profitable opportunities.
  - Telecommunication companies predicting customer LTV
  - Customers understanding pricing of house/apartment
  - Insurance companies identifying fraudulent claims
  - Credit card companies predicting default possibility
  - Retail stores (Target) identifying pregnant customers

$\Rightarrow$ *Regression and Classification*

$\Rightarrow$ *Linear Regression only for the boot camp*

Linear Regression Is an Old Topic

- Linear regression, also called the method of *least squares*, is an old topic, dating back to Gauss in 1795 (he was 18!), later published in this famous book:



THEORIA
MOTVS CORPORVM
COELESTIVM
IN
SECTIONIBVS CONICIS SOLEM AMBIENTIVM
AVCTORE
CAROLO FRIDERICO GAVSS

HAMBVRGI SVMTIBVS FRID. PERTHES ET I. H. BESSER
1809.

Venditur

PARISIIS ap. Treuttel & Würtz. LONDINI ap. R. H. Evans.
STOCKHOLMIAE ap. A. Wiborg. PETROPOLI ap. Klostermann.
MADRITI ap. Sancha. FLORENTIAE ap. Molini, Landi & C°
AMSTELODAMI in libraria: Kunst- und Industrie-Comptoir, dicta.

Linear Regression

- *Simple* Linear Regression
  Regression with *One* Explanatory Variable

- *Multiple* Linear Regression
  Regression with *Multiple* Explanatory Variables

Outline

1. Simple Linear Regression
   - Correlation
   - Least Squares Line
   - R^2
   - Model Diagnostics
   - Inference

2. Multiple Linear Regression
   - Model, Estimate, and Diagnostics
   - Inference
   - Collinearity
   - Categorical Explanatory Variables

Outline

1. Simple Linear Regression
   - Correlation
   - Least Squares Line
   - R^2
   - Model Diagnostics
   - Inference

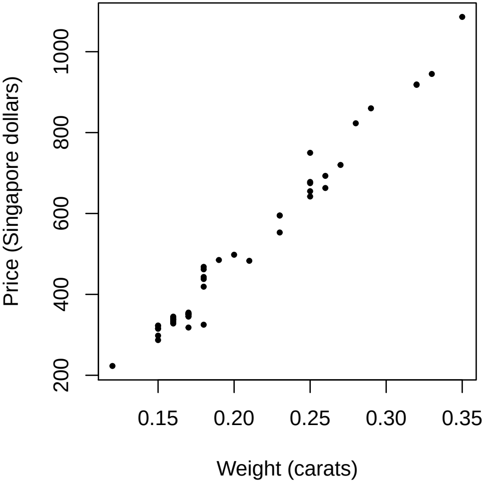
2. Multiple Linear Regression
   - Model, Estimate, and Diagnostics
   - Inference
   - Collinearity
   - Categorical Explanatory Variables

Outline

1. Simple Linear Regression
   - Correlation
   - Least Squares Line
   - R^2
   - Model Diagnostics
   - Inference
2. Multiple Linear Regression
   - Model, Estimate, and Diagnostics
   - Inference
   - Collinearity
   - Categorical Explanatory Variables

An Example of Diamond Ring

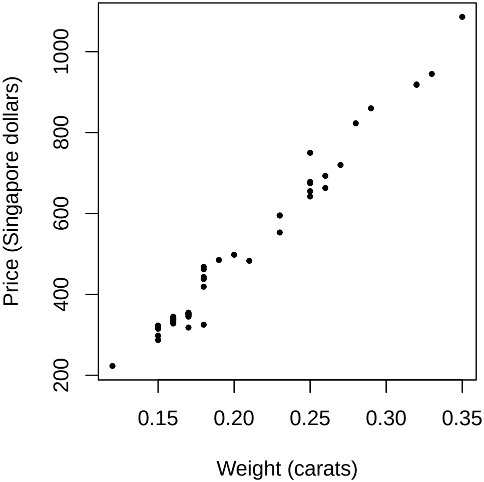
The *scatterplot* below shows the relationship between the price (in Singapore \$) and weight (in carats) of 48 diamond rings.



Response variable?

An Example of Diamond Ring

The *scatterplot* below shows the relationship between the price (in Singapore $) and weight (in carats) of 48 diamond rings.

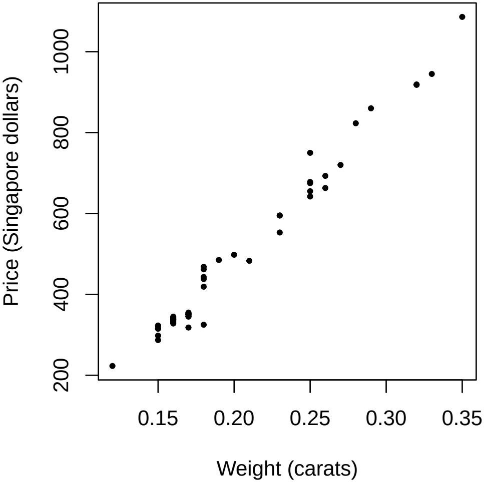


Response variable?

Price

An Example of Diamond Ring

The *scatterplot* below shows the relationship between the price (in Singapore $) and weight (in carats) of 48 diamond rings.



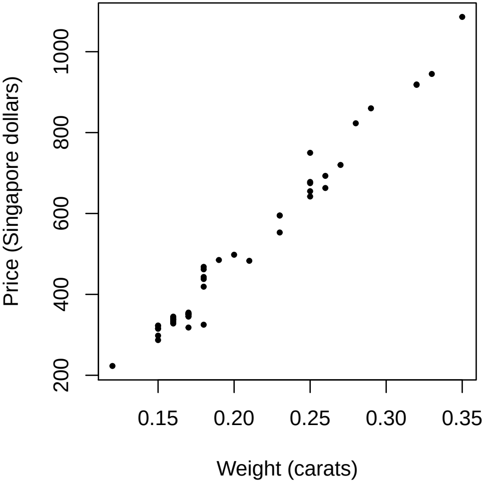
Response variable?

Price

Explanatory variable?

An Example of Diamond Ring

The *scatterplot* below shows the relationship between the price (in Singapore $) and weight (in carats) of 48 diamond rings.



Response variable?

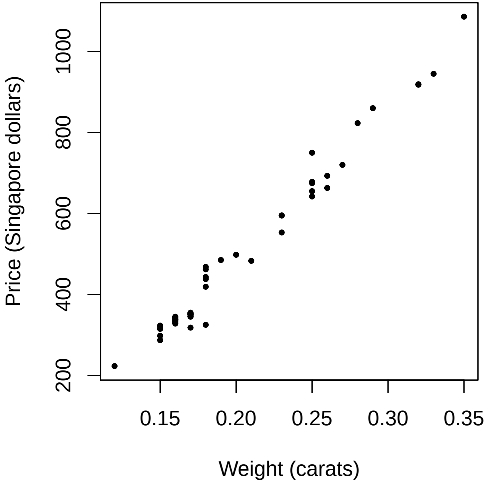
Price

Explanatory variable?

Weight

An Example of Diamond Ring

The *scatterplot* below shows the relationship between the price (in Singapore $) and weight (in carats) of 48 diamond rings.



Response variable?

Price

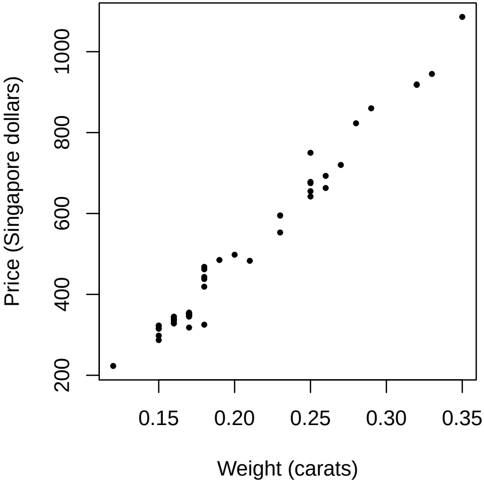
Explanatory variable?

Weight

Relationship? Direction? Strength?

An Example of Diamond Ring

The *scatterplot* below shows the relationship between the price (in Singapore $) and weight (in carats) of 48 diamond rings.



Response variable?

Price

Explanatory variable?

Weight

Relationship? Direction? Strength?

linear, positive, strong

Correlation

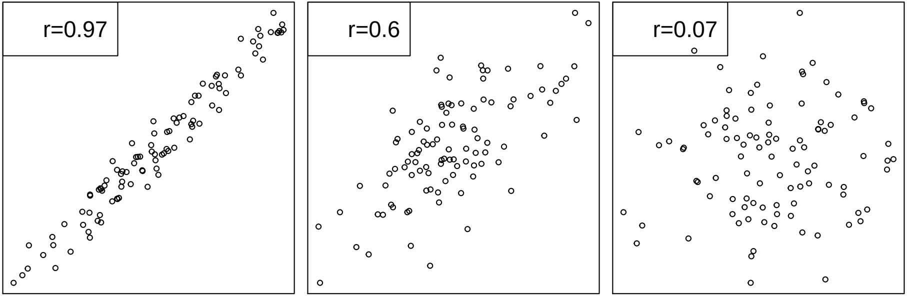
- A quantity that measures the *direction* and *strength* of the linear association between two *quantitative* variables.
- Conventional notation: r

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

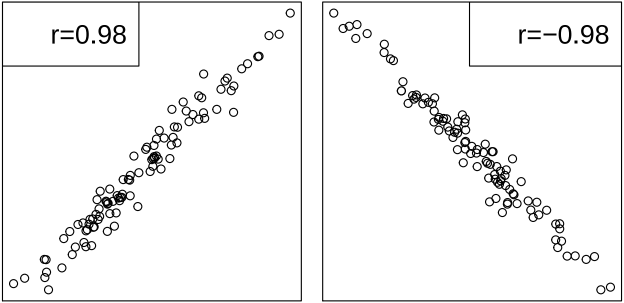
Properties of Correlation (1)

- The *magnitude* (absolute value) of the correlation coefficient measures the *strength* of the linear association between two numerical variables



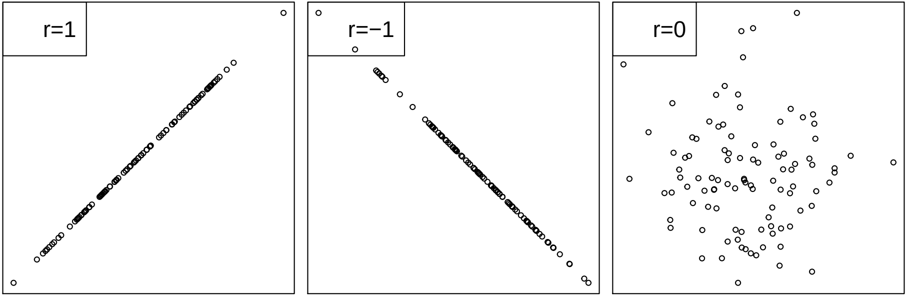
Properties of Correlation (2)

- The *sign* of the correlation coefficient indicates the *direction* of association



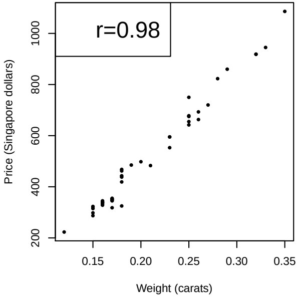
Properties of Correlation (3)

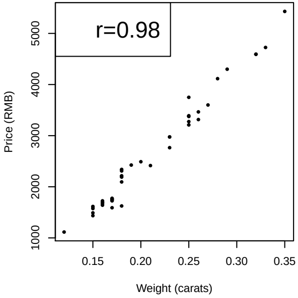
- The correlation coefficient is always between *-1* (*perfect negative* linear association) and *1* (*perfect positive* linear association)
- *0* indicates *no* linear relationship

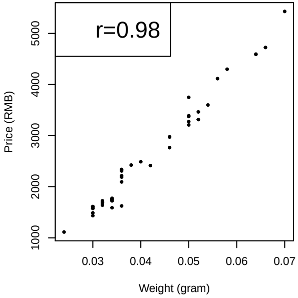


Properties of Correlation (4)

- The correlation coefficient is *unitless*, and is not affected by changes in the center or scale of either variable (such as unit conversions)

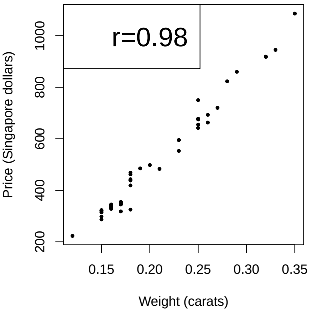


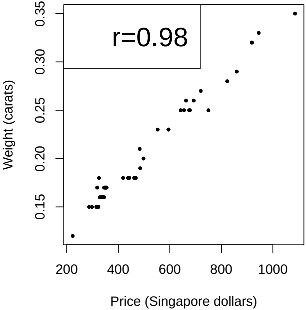




Properties of Correlation (5)

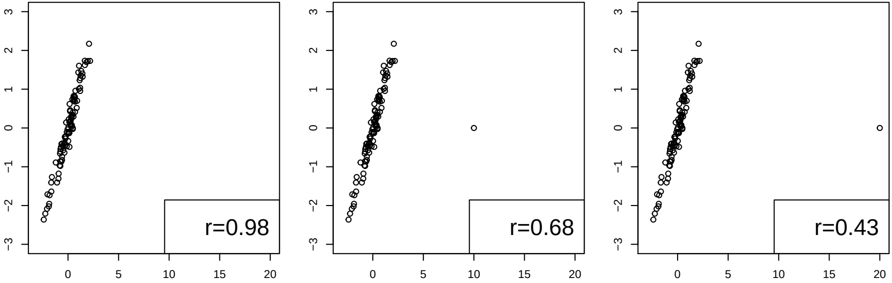
- The correlation of X with Y is the *same* as of Y with X





Properties of Correlation (6)

- The correlation coefficient is *sensitive to outliers*

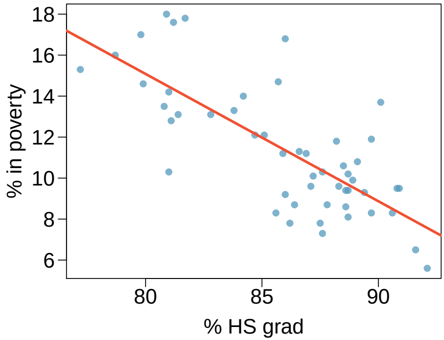


Guessing the Correlation

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line

Which of the following is the best guess for the correlation?

a. 0.6
b. -0.75
c. -0.1
d. 0.02
e. -1.5

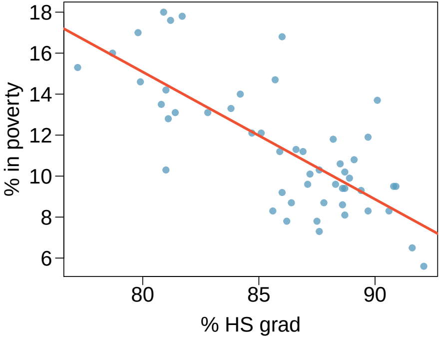


Guessing the Correlation

The scatterplot below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line

Which of the following is the best guess for the correlation?

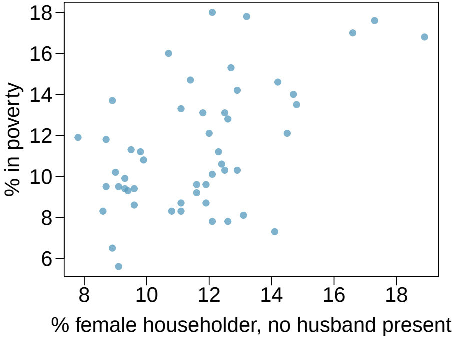
(a) 0.6
(b) *-0.75*
(c) -0.1
(d) 0.02
(e) -1.5



Guessing the Correlation

Which of the following is the best guess for the correlation between % in poverty and % female householder?

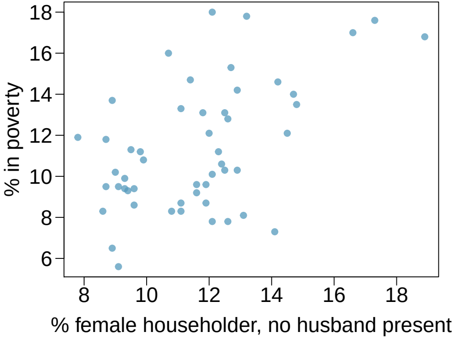
(a) 0.1
(b) -0.6
(c) -0.4
(d) 0.9
(e) 0.5



Guessing the Correlation

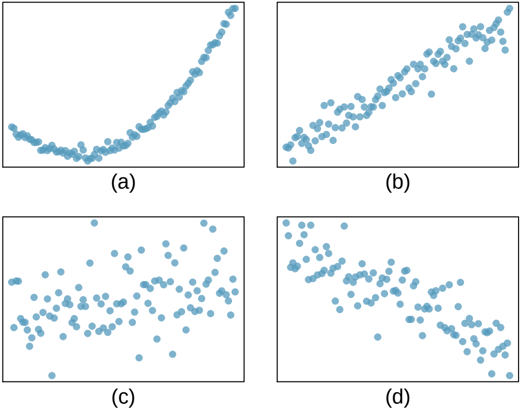
Which of the following is the best guess for the correlation between % in poverty and % female householder?

1. 0.1
2. -0.6
3. -0.4
4. 0.9
5. *0.5*



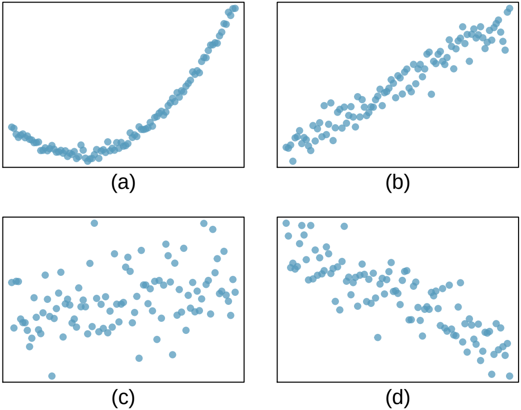
Guessing the Correlation

Which of the following has the strongest correlation, i.e. correlation coefficient closest to $+1$ or -1?



Guessing the Correlation

Which of the following has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?

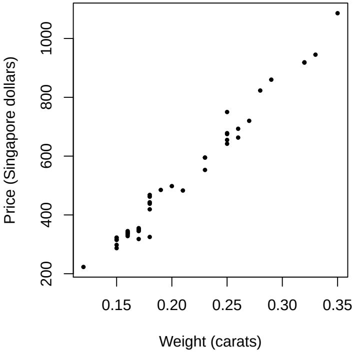


(b) $\rightarrow$ correlation means linear association

Outline

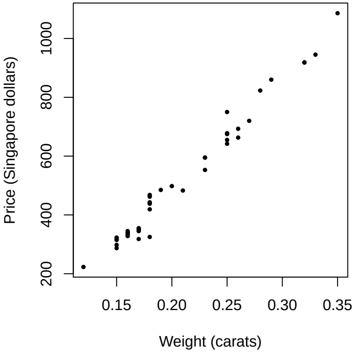
Simple Linear Relationship

- How do weight and price vary together?



Simple Linear Relationship

- How do weight and price vary together?



- How much do I *expect to pay* for a ring with a 0.3 carat diamond?
- Given a budget of \$1000, *how big can I expect to afford*?

The Simple Linear Regression Setup

- Observe n independent pairs $(x_1, y_1), \ldots, (x_n, y_n)$ where x explains/predicts y
  - x is called the independent variable, explanatory variable or predictor
  - y is called the dependent variable or response

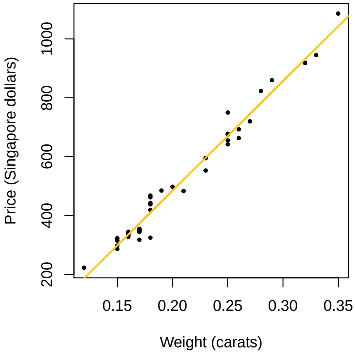
The Simple Linear Regression Setup

- Observe n independent pairs $(x_1, y_1), \ldots, (x_n, y_n)$ where x explains/predicts y
  - x is called the independent variable, explanatory variable or predictor
  - y is called the dependent variable or response
- Examples
  - Weight of diamond vs price
  - Market return vs stock return
  - Education vs salary

The Simple Linear Regression Setup

- Observe n independent pairs $(x_1, y_1), ..., (x_n, y_n)$ where x explains/predicts y
  - x is called the independent variable, explanatory variable or predictor
  - y is called the dependent variable or response
- Examples
  - Weight of diamond vs price
  - Market return vs stock return
  - Education vs salary
- Simple linear regression line
  - Summarizes the *linear* relationship between x and y.
  - Describes how y *changes* as x changes.
  - Is often used as a mathematical model to use x to *predict* y.

Quantify Linear Relationship



$$\hat{price}_i = b_0 + b_1 weight_i$$

predicted price

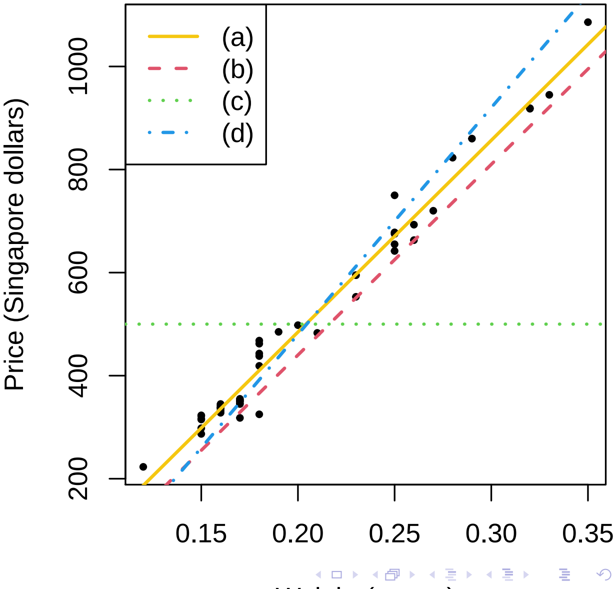
intercept

slope

explanatory variable weight

Eyeballing the line

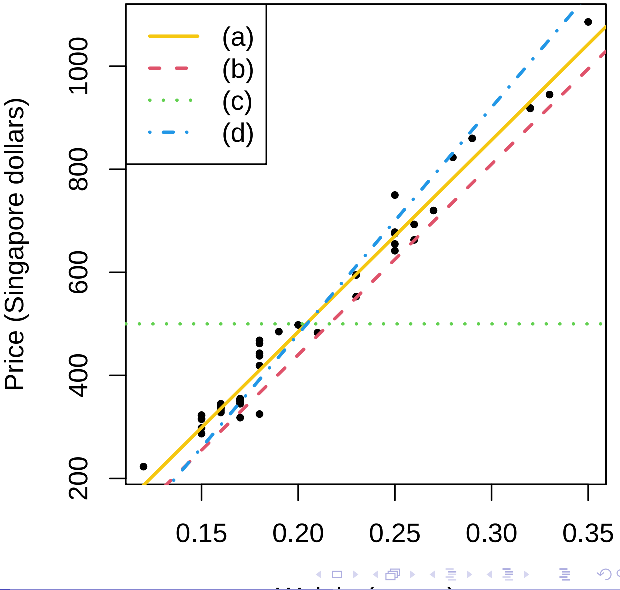
Which of the following appears to be the line that best fits the linear relationship between price and weight? Choose one.



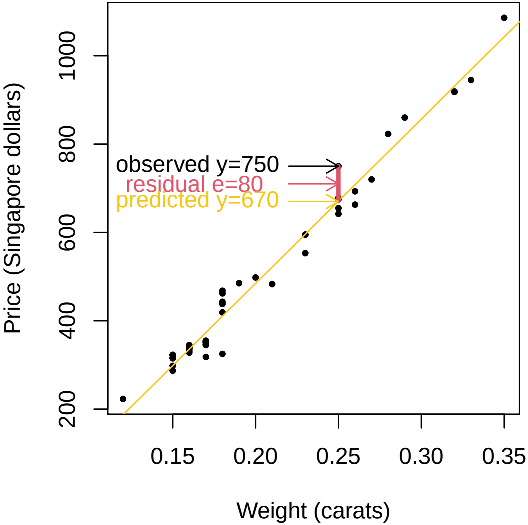
Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between price and weight? Choose one.

(a)



Choosing the Line



The price for this ring is 80 dollars more than predicted.

The Least Squares Line

$$\hat{y}_i = b_0 + b_1 x_i$$

predicted y *intercept* *slope* *explanatory variable*

- For each observation in the data set, your *line* predicts where y should be.
- The *residual* from ith data point is how far the true y value is from where the line predicts.

$$e_i = y_i - \hat{y}_i$$

The Least Squares Line

$$\hat{y}_i = b_0 + b_1 x_i$$

predicted y

intercept

slope

explanatory variable

- For each observation in the data set, your *line* predicts where y should be.
- The *residual* from ith data point is how far the true y value is from where the line predicts.

$$e_i = y_i - \hat{y}_i$$

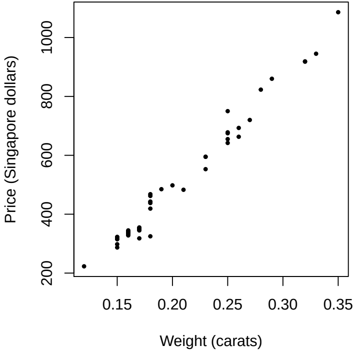
- Least-squares criterion: Find b_0, b_1 to minimize the residual sum of squares (RSS) or sum of squared errors (SSE)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Given...

```
> summarizeColumns(diamond) %>%
+   kable(digits = 2)
```

```
|name    |type     | na|   mean|   disp| median|    mad|    min|     max| nlevs|
|:-------|:--------|--:|------:|------:|------:|------:|------:|-------:|-----:|
|weight  |numeric  |  0|   0.20|   0.06|   0.18|   0.04|   0.12|    0.35|     0|
|price   |integer  |  0| 500.08| 213.64| 428.50| 157.16| 223.00| 1086.00|     0|
```



	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

Slope

The slope of the regression can be calculated as

$$b_1 = r \times \frac{s_y}{s_x}$$

Slope

The slope of the regression can be calculated as

$$b_1 = r \times \frac{s_y}{s_x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$b_1 = ...?$

Slope

The slope of the regression can be calculated as

$$b_1 = r \times \frac{s_y}{s_x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$$b_1 = 0.98 \times \frac{213.64}{0.06} = 3721.02$$

Slope

The slope of the regression can be calculated as

$$b_1 = r \times \frac{s_y}{s_x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$$b_1 = 0.98 \times \frac{213.64}{0.06} = 3721.02$$

Interpretation

Slope

The slope of the regression can be calculated as

$$b_1 = r \times \frac{s_y}{s_x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$$b_1 = 0.98 \times \frac{213.64}{0.06} = 3721.02$$

Interpretation

each 1 carat increase in weight results in SGD$ 3,721 increase in price

each 0.1 carat increase in weight results in SGD$ 372.1 increase in price

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1\bar{x}$$

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$$b_0 = 500.08 - 3721.02 \times 0.20 = -259.63$$

Intercept

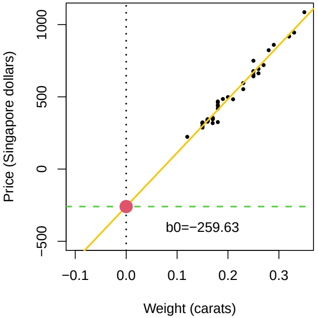
The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1\bar{x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$$b_0 = 500.08 - 3721.02 \times 0.20 = -259.63$$



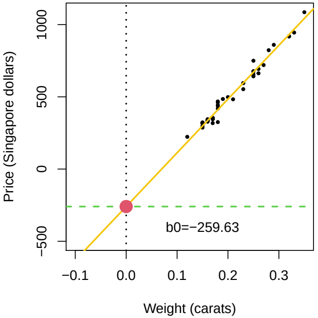
Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	



$$b_0 = 500.08 - 3721.02 \times 0.20 = -259.63$$

Interpretation

Intercept

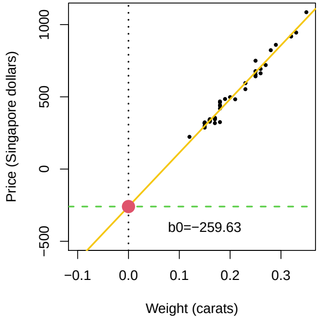
The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1 \bar{x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	

$$b_0 = 500.08 - 3721.02 \times 0.20 = -259.63$$



Interpretation

If you walk into the store, asking for a 0-carat ring, the store actually pays you SGD$259.63!

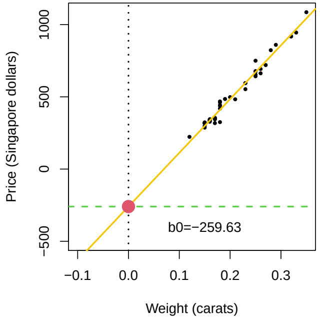
Intercept

The intercept is where the regression line intersects the y-axis. The calculation of the intercept uses the fact the a regression line always passes through $(\bar{x}, \bar{y})$.

$$b_0 = \bar{y} - b_1\bar{x}$$

In context...

	weight (x)	price (y)
mean	$\bar{x} = 0.20$	$\bar{y} = 500.08$
sd	$s_x = 0.06$	$s_y = 213.64$
correlation	$r = 0.98$	



$$b_0 = 500.08 - 3721.02 \times 0.20 = -259.63$$

Interpretation

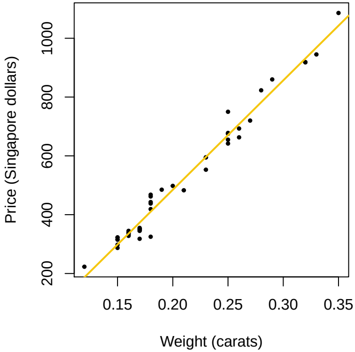
If you walk into the store, asking for a 0-carat ring, the store actually pays you SGD$259.63!

Don't extrapolate outside the data range!

Regression Line

$$\widehat{price} = -259.63 + 3721.02 \ weight$$

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$
(Intercept)	-259.63	17.32	-14.99	$<$2e-16 ***
weight	3721.02	81.79	45.50	$<$2e-16 ***



Summary

- Coefficients $b_1 = r \times \frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$.
- Predicted value $\hat{y}_i = b_0 + b_1 x_i$.
- Actual value y_i.
- Residual $e_i = y_i - \hat{y}_i$.
- We choose the line to make SSE or RSS as small as possible.

Correlation and Regression Line

- Both for linear relationship between two variables.
  - Same sign between b_1 and r.

Correlation and Regression Line

- Both for linear relationship between two variables.
  - Same sign between b_1 and r.
- r does not depend on which is x and which is y.

Correlation and Regression Line

- Both for linear relationship between two variables.
  - Same sign between b_1 and r.
- r does not depend on which is x and which is y.
- But the regression line does.

$$
\begin{aligned}
\widehat{price} &= b_0 + b_1 \ weight \\
\widehat{weight} &= a_0 + a_1 \ price
\end{aligned}
$$

Correlation and Regression Line

- Both for linear relationship between two variables.
  - Same sign between b_1 and r.
- r does not depend on which is x and which is y.
- But the regression line does.

$$\begin{aligned} \widehat{price} &= b_0 + b_1 \ weight \\ \widehat{weight} &= a_0 + a_1 \ price \end{aligned}$$

- What is the relationship between b_1 and a_1?

Correlation and Regression Line

- Both for linear relationship between two variables.
  - Same sign between b_1 and r.
- r does not depend on which is x and which is y.
- But the regression line does.

$$\begin{aligned}\widehat{price} &= b_0 + b_1 \; weight \\ \widehat{weight} &= a_0 + a_1 \; price\end{aligned}$$

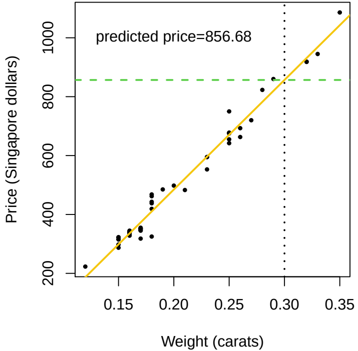
- What is the relationship between b_1 and a_1?

  $b_1 = r \times \frac{s_y}{s_x}$, $a_1 = r \times \frac{s_x}{s_y}$.

  So (1) $b_1 \times a_1 = r^2 \in [0, 1]$, *(2) If* $s_x = s_y$, $b_1 = a_1$.

Prediction

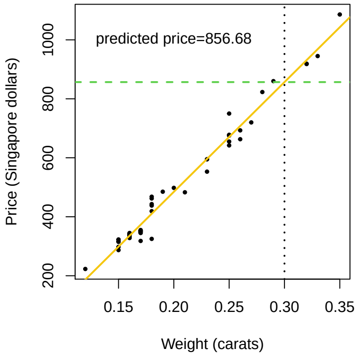
- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- According to the linear model $price = -259.63 + 3721.02\ weight$, what is the predicted price for a 0.30 carat ring?



Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- According to the linear model $price = -259.63 + 3721.02\ weight$, what is the predicted price for a 0.30 carat ring?

  $-259.63 + 3721.02 \times 0.30 = 856.68.$



Outline

How Good is the Regression Model?

- To see how accurate we can predict Y, we can look at the standard deviation of the residuals.
- *Root Mean Squared Error (RMSE)*:

$$MSE = \frac{SSE}{n-2} \qquad RMSE = \sqrt{MSE}$$

How Good is the Regression Model?

- To see how accurate we can predict Y, we can look at the standard deviation of the residuals.
- *Root Mean Squared Error (RMSE)*:

$$MSE = \frac{SSE}{n-2} \qquad RMSE = \sqrt{MSE}$$

- RMSE tells us how far our predictions are off on average. (RMSE=\$31.84 in diamond example)
- If RMSE is small (close to zero), then the regression is doing a good job.

How Good is the Regression Model?

- To see how accurate we can predict Y, we can look at the standard deviation of the residuals.
- *Root Mean Squared Error (RMSE)*:

$$MSE = \frac{SSE}{n-2} \qquad RMSE = \sqrt{MSE}$$

- RMSE tells us how far our predictions are off on average. (RMSE=$31.84 in diamond example)
- If RMSE is small (close to zero), then the regression is doing a good job.
- *Drawback*: RMSE depends on the size of Y.
  Example: diamond price in RMB vs Singapore$

RMSE

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                                         RMSE
Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

R-Square: R^2

- Some of the variation in Y can be explained by variation in X and some cannot.
  - Diamond ring, price variation:
    weight variation + purity variation.
- R^2: fraction of variance that can be explained by X.

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.

R-Square: R^2

- Some of the variation in Y can be explained by variation in X and some cannot.
  - Diamond ring, price variation:
    weight variation + purity variation.
- R^2: fraction of variance that can be explained by X.

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

  where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.
- Fact: $R^2 = r^2$ with r being the correlation.

R-Square: R^2

- Some of the variation in Y can be explained by variation in X and some cannot.
  - Diamond ring, price variation:
    weight variation + purity variation.
- R^2: fraction of variance that can be explained by X.

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS}$$

  where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.
- Fact: $R^2 = r^2$ with r being the correlation.
- The larger R^2, the stronger the *linear relationship*; the more confident we are in our prediction.

R^2

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max
-85.159 -21.448  -0.869  18.972  79.370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783     Adjusted R-squared:  0.9778
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

R Square

More about R^2

- *Interpretation*: "the proportion of variation explained by the regression", where the variation refers to sample variance of y
- R^2 captures the usefulness of using x to predict y, and is often used as a measure of the "effectiveness" of a regression.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.

More about R^2

- *Interpretation*: "the proportion of variation explained by the regression", where the variation refers to sample variance of y
- R^2 captures the usefulness of using x to predict y, and is often used as a measure of the "effectiveness" of a regression.
- The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.
- *R^2 is not useful for deciding between regressions when*
  - The response variables y are different due to transformation.
  - The data points are different due to the removal of outliers.

Outline

1. Simple Linear Regression
   - Correlation
   - Least Squares Line
   - R^2
   - Model Diagnostics
   - Inference

2. Multiple Linear Regression
   - Model, Estimate, and Diagnostics
   - Inference
   - Collinearity
   - Categorical Explanatory Variables

The Simple Linear Regression Model

- The data $(x_1, y_1), \ldots, (x_n, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i IID $\sim N(0, \sigma^2)$

The Simple Linear Regression Model

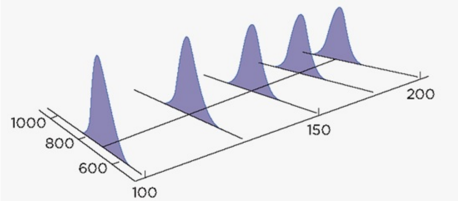
- The data $(x_1, y_1), \ldots, (x_n, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i IID $\sim N(0, \sigma^2)$

- The average values of the response fall on a line

$$E(y_i | x_i) = \beta_0 + \beta_1 x_i$$



Parameters of the Model

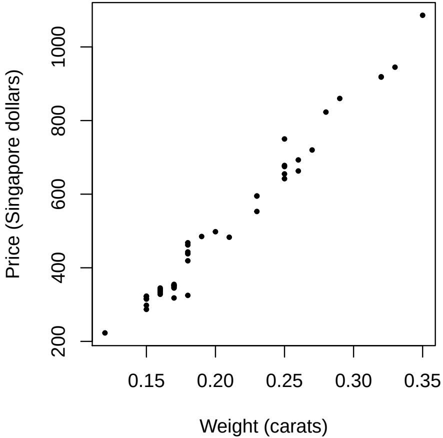
- Model parameters: $\beta_0, \beta_1, \sigma^2$
- Think of data as the sum of the *signal* $\beta_0 + \beta_1 x_i$ and *noise* ϵ_i.
- *Interpretations:*
    - $\beta_0 + \beta_1 x_i$: conditional mean of y at $x = x_i$.
    - β_0: average of y when $x = 0$.
    - β_1: average change in y between locations x and $x + 1$.
    - σ: standard deviation of variation around conditional mean.
- *Properties of the errors:*
    - Independence
    - Equal variance
    - Normally distributed

Model Diagnostics

- Make sure there are no gross violations of the model
  - Is the relationship between x and y *linear*?
  - Do the residuals show iid normal behavior (i.e., *independent, equal variance, normality*)?
  - Are there *outliers* that may distort the model fit?
- Three crucial steps in checking a model
  - A *y vs. x scatterplot* should reveal a linear pattern, linear dependence.
  - A *Residual vs. x scatterplot* should reveal no meaningful pattern.
  - A *Residual vs. Predicted scatterplot* should reveal no meaningful pattern.
  - A *histogram and normal quantile plot* of the residuals should be consistent with the assumption of normality of the errors.

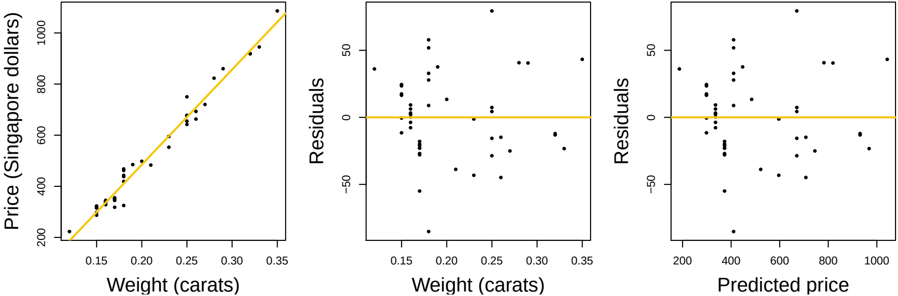
Model Diagnostics: Diamond Ring

- The scatterplot of Price vs. Weight clearly indicates a linear relationship

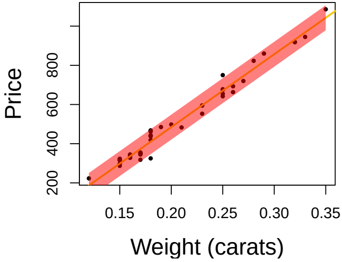


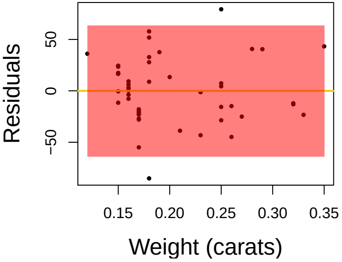
Model Diagnostics: Diamond Ring

- These plots reveal no systematic pattern in the residuals, which is good.



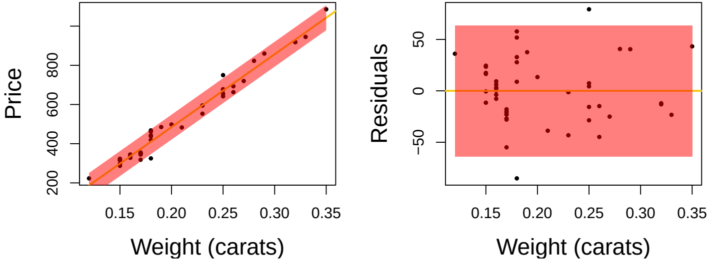
Model Diagnostics: Diamond Ring





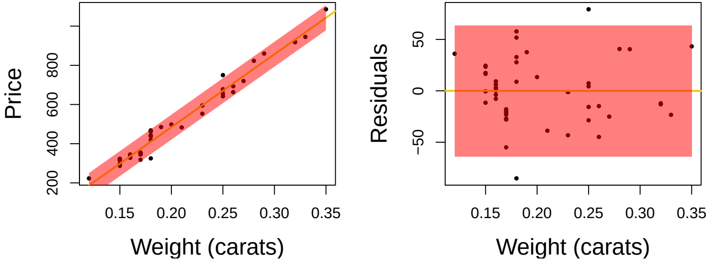
- The variability of points around the least squares line should be roughly constant.

Model Diagnostics: Diamond Ring



- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the fitted LS line should be roughly constant as well.

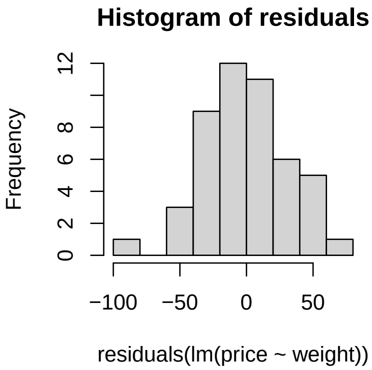
Model Diagnostics: Diamond Ring

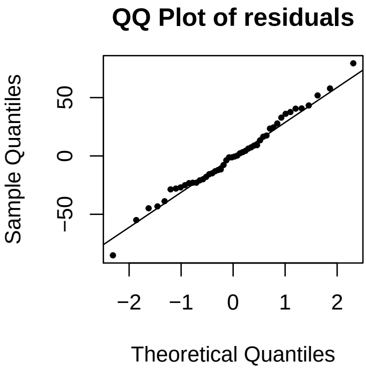


- The variability of points around the least squares line should be roughly constant.
- This implies that the variability of residuals around the fitted LS line should be roughly constant as well.
- Also called *homoscedasticity*. Otherwise called *heteroscedasticity*.

Model Diagnostics: Diamond Ring

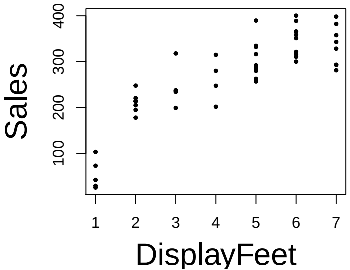
- The residual histogram is consistent with the normality assumption.
- QQplot: The points stay close to the line, suggesting normality.



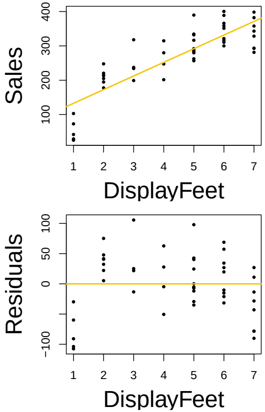


Model Diagnostics 1: Non-linearity

- A chain of liquor stores needs to know how much shelf space in its stores to devote to showing a new wine to maximize its profit.
- Space devoted to other products brings in about \$50 of net revenue per linear foot.
- The data has sales (\$) and shelf-feet per week from 47 stores of the chain.
- Should we expect a *linear* relationship between promotion and sales, or should we expect *diminishing* marginal gains?

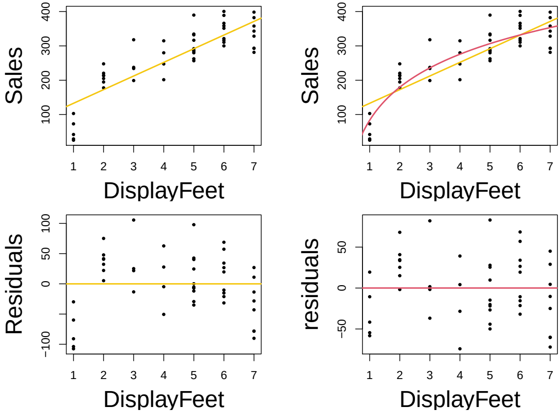


Model Diagnostics 1: A Linear Fit Does Not Make Sense!



- The line misses important features in the data.
- More obvious in residual plot.

Model Diagnostics 1: A Linear Fit Does Not Make Sense!



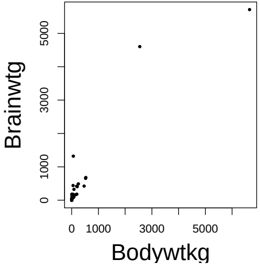
- The line misses important features in the data.
- More obvious in residual plot.
- $\log(x)$ *transformation*

Model Diagnostics 1: Non-linear

```
> summary(lm(Sales~DisplayFeet))

Call:
lm(formula = Sales ~ DisplayFeet)

Residuals:
     Min       1Q   Median       3Q      Max
-107.489  -29.552    0.085   33.342  105.598

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    93.03      18.23   5.104 6.50e-06 ***
DisplayFeet    39.76       3.77  10.547 9.55e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.59 on 45 degrees of freedom
Multiple R-squared:  0.712,     Adjusted R-squared:  0.7056
F-statistic: 111.2 on 1 and 45 DF,  p-value: 9.555e-14

> summary(lm(Sales~logDisplayFeet))

Call:
lm(formula = Sales ~ logDisplayFeet)

Residuals:
    Min      1Q  Median      3Q     Max
-74.230 -27.596  -1.751  28.417  83.038

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      83.560     14.413   5.797 6.24e-07 ***
logDisplayFeet  138.621      9.834  14.096  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.31 on 45 degrees of freedom
Multiple R-squared:  0.8153,    Adjusted R-squared:  0.8112
F-statistic: 198.7 on 1 and 45 DF,  p-value: < 2.2e-16
```

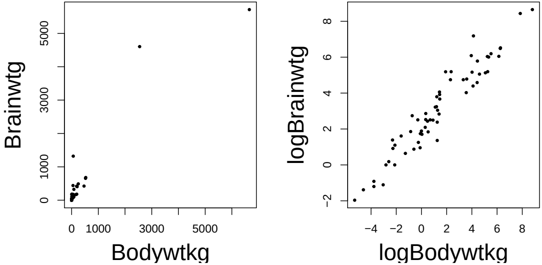
Model Diagnostics 1: Log-log transformation

- Mammals.dat describes 62 terrestrial mammals, with body weight in kilograms and brain weight in grams.



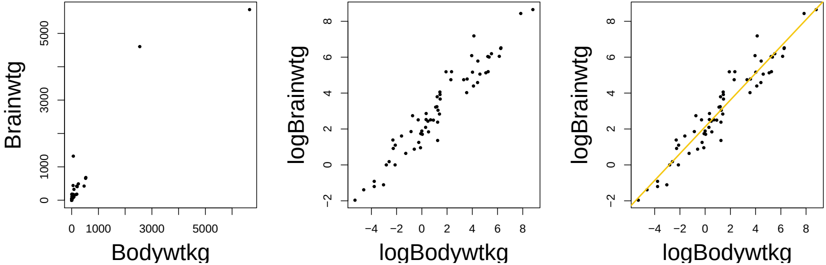
Model Diagnostics 1: Log-log transformation

- Mammals.dat describes 62 terrestrial mammals, with body weight in kilograms and brain weight in grams.



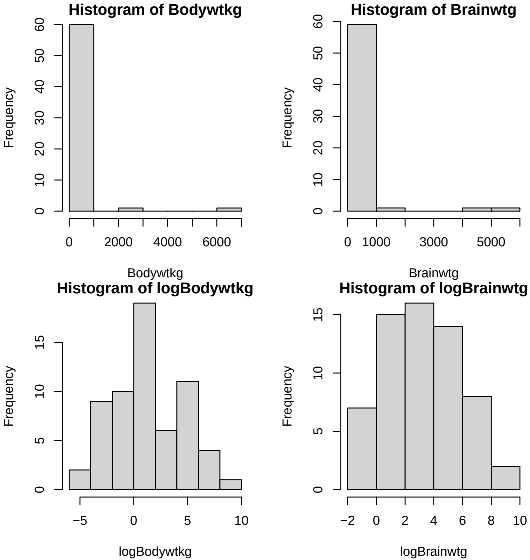
Model Diagnostics 1: Log-log transformation

- Mammals.dat describes 62 terrestrial mammals, with body weight in kilograms and brain weight in grams.



- $logBrainwtg = 2.12 + 0.75logBodywtkg$
- Multiplicative relationship, instead of additive
- For 1% increase in log Body Wt, log Brain Wt increases 0.75%.

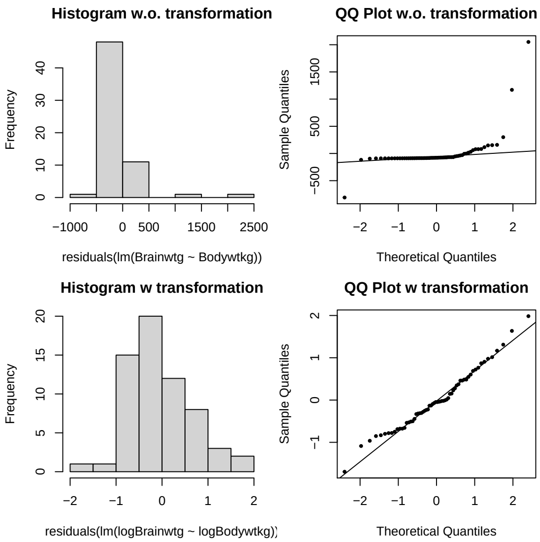
Model Diagnostics 1: Log-log transformation



Interpretation for Models on Log-scale

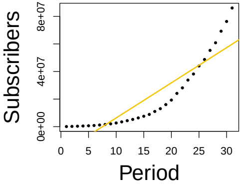
- $y = a + bx$
  If x changes from x to $x + \delta$, the exact change in y is $b\delta$.
- $y = a + b \log_e x$
  If x changes from x to $x(1 + p\%)$ (a $p\%$ change), the approximate change in y is $bp\%$.
- $\log_e y = a + bx$
  If x changes from x to $x + \delta$, the approximate change in y is y to $y(1 + b\delta)$.
- $\log_e y = a + b \log_e x$
  If x changes from x to $x(1 + p\%)$ (a $p\%$ change), the approximate change in y is y to $y(1 + bp\%)$.

Model Diagnostics 2: Non-normal Residuals



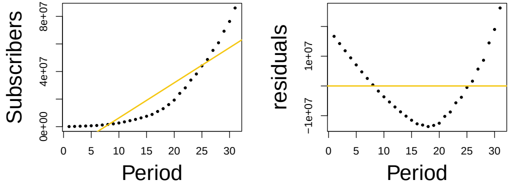
Model Diagnostics 3: Auto-correlated Residuals

- The file cellular.dat contains the number of subscribers to cell phone service in the US every six months from the end of 1984 to the end of 1995.



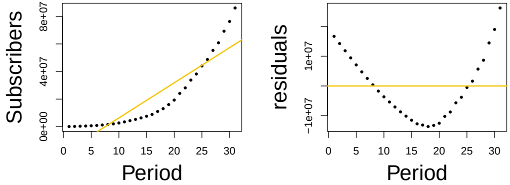
Model Diagnostics 3: Auto-correlated Residuals

- The file cellular.dat contains the number of subscribers to cell phone service in the US every six months from the end of 1984 to the end of 1995.



Model Diagnostics 3: Auto-correlated Residuals

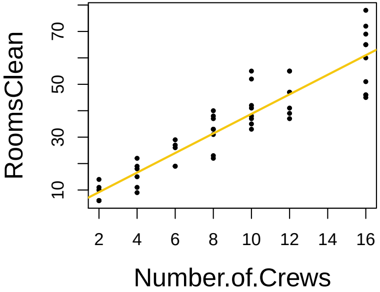
- The file cellular.dat contains the number of subscribers to cell phone service in the US every six months from the end of 1984 to the end of 1995.



- *Meandering pattern* shows that the residuals violate the independence assumption, i.e. *auto-correlated*

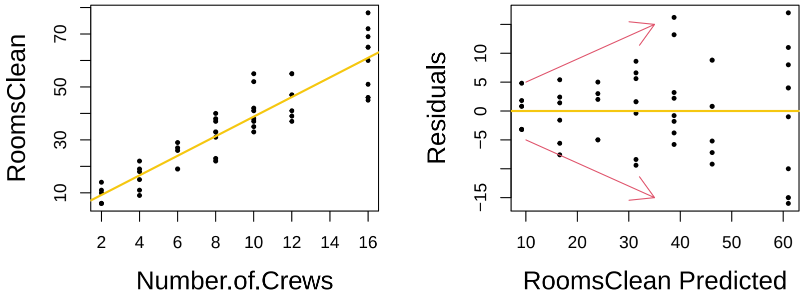
Model Diagnostics 4: Non-constant Variability or Heteroscedasticity

- The file cleaning.dat contains the number of crews and the # of rooms cleaned for 53 teams of building maintenance workers.



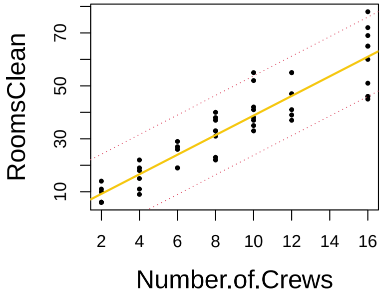
Model Diagnostics 4: Non-constant Variability or Heteroscedasticity

- The file cleaning.dat contains the number of crews and the # of rooms cleaned for 53 teams of building maintenance workers.



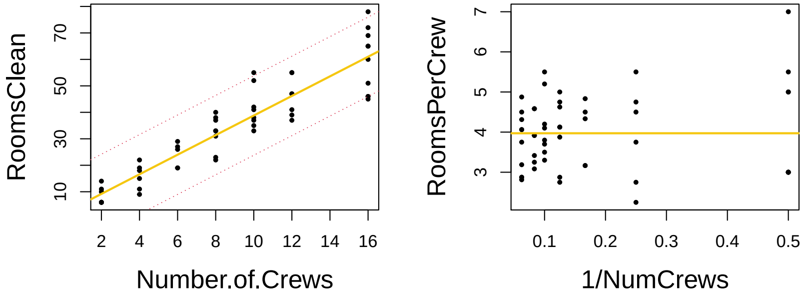
- Because the residuals *fan out* as the # of crews increases, these data *violate* the assumption of *equal error variance* in the model.

Model Diagnostics 4: Non-constant Variability or Heteroscedasticity



- Over and under estimate variance for different x region

Model Diagnostics 4: Non-constant Variability or Heteroscedasticity



- Over and under estimate variance for different x region
- Consider RoomsClean per Crew as y

Model Diagnostics 5: Outliers

- *Outliers* are points that lie away from the cloud of points.

Model Diagnostics 5: Outliers

- *Outliers* are points that lie away from the cloud of points.
- *High leverage* points: outliers that lie horizontally away from the center of the cloud.

Model Diagnostics 5: Outliers

- *Outliers* are points that lie away from the cloud of points.
- *High leverage* points: outliers that lie horizontally away from the center of the cloud.
- *Influential* points: high leverage points that actually influence the slope of the regression line.

Model Diagnostics 5: Outliers

- *Outliers* are points that lie away from the cloud of points.
- *High leverage* points: outliers that lie horizontally away from the center of the cloud.
- *Influential* points: high leverage points that actually influence the slope of the regression line.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it's not an influential point.

Leverage

- Leverage is a property of the values of the *predictor*, not the response.

Leverage

- Leverage is a property of the values of the *predictor*, not the response.
- Leverage points are *often* influential in determining the fitted line.
  - Analogy: seesaw, your weight is more influential at the end of a see-saw compared to near to the middle.

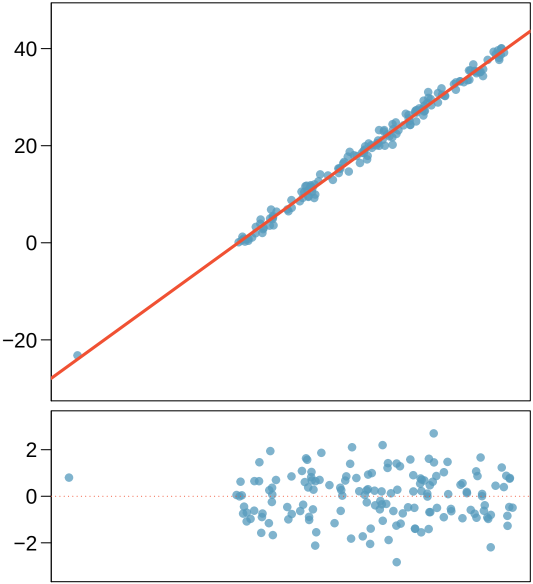
Leverage

- Leverage is a property of the values of the *predictor*, not the response.
- Leverage points are *often* influential in determining the fitted line.
  - Analogy: seesaw, your weight is more influential at the end of a see-saw compared to near to the middle.
- Leverage points are *not necessarily bad*, and can improve the precision of the slope estimate.
  - Imagine a point follows the line but far away from the rest of the points

Model Diagnostics 5: Outliers

Which of the below best describes the outlier?

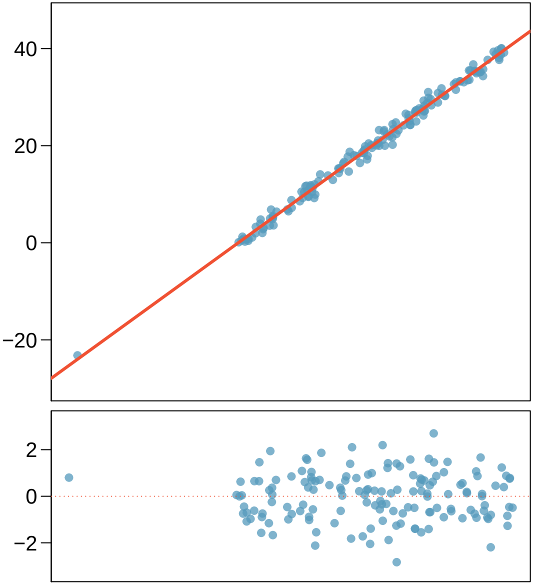
(a) influential
(b) high leverage
(c) none of the above
(d) there are no outliers



Model Diagnostics 5: Outliers

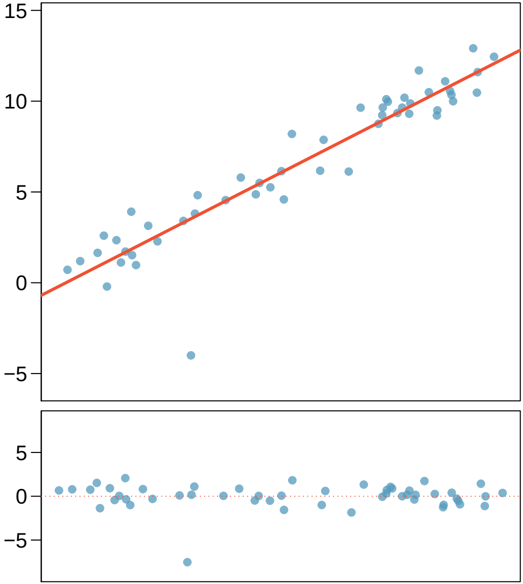
Which of the below best describes the outlier?

- (a) influential
- (b) *high leverage*
- (c) none of the above
- (d) there are no outliers



Model Diagnostics 5: Outliers

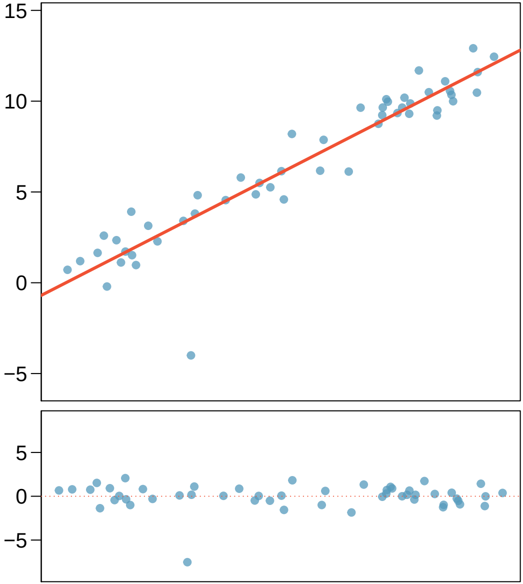
Does this outlier influence the slope of the regression line?



Model Diagnostics 5: Outliers

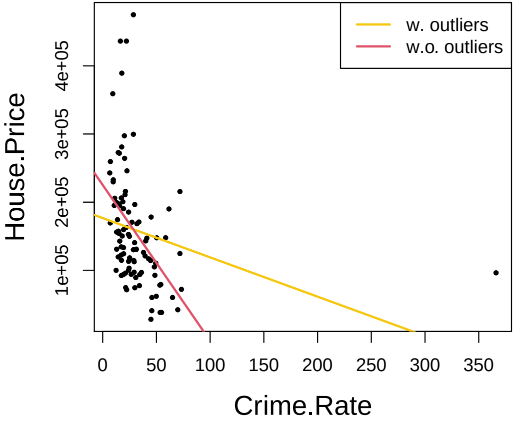
Does this outlier influence the slope of the regression line?

Not much...



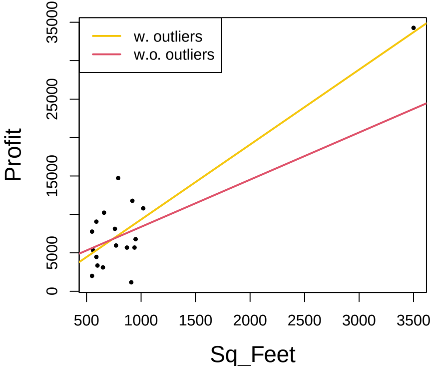
Model Diagnostics 5: Outliers

- The file phila.dat contains average prices of houses sold and crime rates for 110 communities in/near Philadelphia in April 1996.



Model Diagnostics 5: Outliers

- Leverage points can impact inferences in dramatic fashion.
- The data cottages.dat contains the profits obtained by a construction firm for 18 properties, as well as the square footage of each of the properties.
- Based on this data, should the firm continue to build large properties?



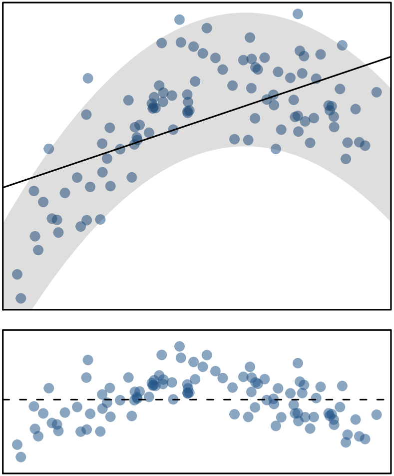
Model Diagnostics 5: Outliers

```
> summary(lm(Profit~Sq_Feet))

Call:
lm(formula = Profit ~ Sq_Feet)

Residuals:
    Min      1Q  Median      3Q     Max 
-7288.8 -2307.2  -289.6  2428.9  7442.9 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -416.859   1437.015  -0.290    0.775    
Sq_Feet        9.751      1.296   7.524 1.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3570 on 16 degrees of freedom
Multiple R-squared:  0.7797,    Adjusted R-squared:  0.7659 
F-statistic: 56.62 on 1 and 16 DF,  p-value: 1.217e-06

> summary(lm(Profit~Sq_Feet,subset = Sq_Feet < max(Sq_Feet,na.rm = TRUE)))

Call:
lm(formula = Profit ~ Sq_Feet, subset = Sq_Feet < max(Sq_Feet, 
    na.rm = TRUE))

Residuals:
   Min     1Q Median     3Q    Max 
 -6663  -2327  -1015   2288   7635 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2245.400   4237.249   0.530    0.604
Sq_Feet        6.137      5.557   1.104    0.287

Residual standard error: 3634 on 15 degrees of freedom
Multiple R-squared:  0.07521,   Adjusted R-squared:  0.01355 
F-statistic:  1.22 on 1 and 15 DF,  p-value: 0.2868
```

- Which version of this model should the firm use to estimate profits on the next large cottage it is considering building?

Model Diagnostics: Checking Conditions

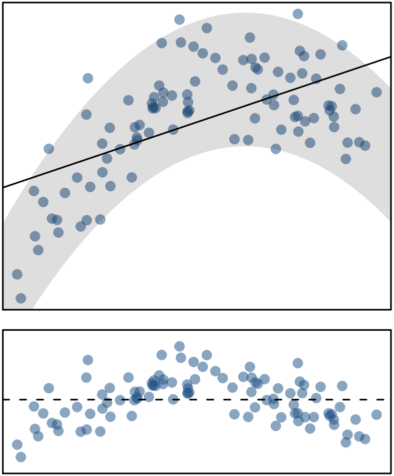
What condition is this linear model obviously violating?



(a) Constant variability
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers

Model Diagnostics: Checking Conditions

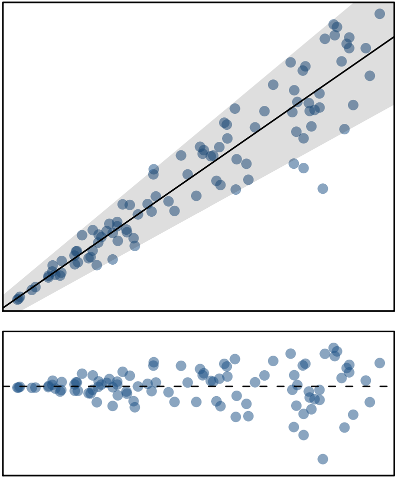
What condition is this linear model obviously violating?



- (a) Constant variability
- (b) *Linear relationship*
- (c) Normal residuals
- (d) No extreme outliers

Model Diagnostics: Checking Conditions

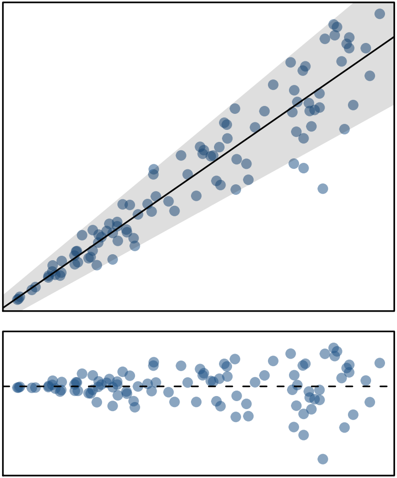
What condition is this linear model obviously violating?



- (a) Constant variability
- (b) Linear relationship
- (c) Normal residuals
- (d) No extreme outliers

Model Diagnostics: Checking Conditions

What condition is this linear model obviously violating?



(a) *Constant variability*
(b) Linear relationship
(c) Normal residuals
(d) No extreme outliers

Outline

Recall of the Simple Linear Regression Model

- To perform statistical inference, use simple linear regression model.
- The data $(x_1, y_1), ..., (x_n, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i iid $\sim N(0, \sigma^2)$

Recall of the Simple Linear Regression Model

- To perform statistical inference, use simple linear regression model.
- The data $(x_1, y_1), \ldots, (x_n, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i iid $\sim N(0, \sigma^2)$
- Model parameters, $\beta_0, \beta_2, \sigma^2$

Estimate of the Regression Line

- Population line

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Least squares line

$$y_i = b_0 + b_1 x_i + e_i$$

Estimate of the Regression Line

- Population line

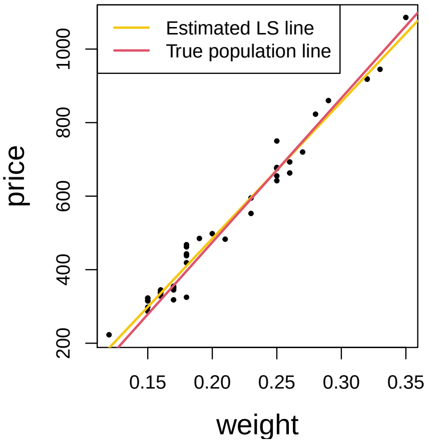
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Least squares line

$$y_i = b_0 + b_1 x_i + e_i$$

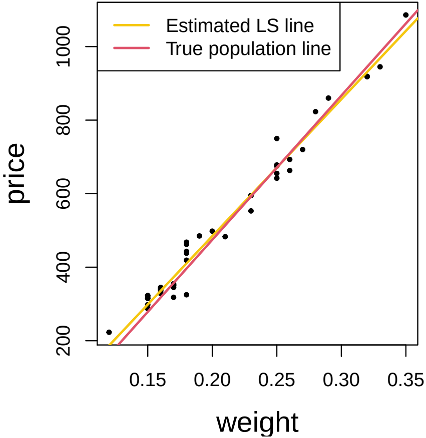
- Use *RMSE* to estimate σ

Inference about the Regression Line



- The regression line from the sample is not the regression line from the population.

Inference about the Regression Line



- The regression line from the sample is not the regression line from the population.
- We want to
  - assess how well the line describes the relationship.
  - guess the slope of the population line.
  - guess what value Y would take for a given X value

Sampling Distributions

- b_1 is normal with mean β_1 and its SE
- b_0 is normal with mean β_0 and its SE

Relevant Questions

1. Is $\beta_1 = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
   - We can address this using hypothesis testing.
   - $H_0 : \beta_1 = 0$ (nothing is going on): The explanatory variable is not a significant predictor of the response variable, i.e. no relationship $\Rightarrow$ slope of the relationship is 0.
   - $H_a : \beta_1 \neq 0$ (something going on): The explanatory variable is a significant predictor of the response variable, i.e. relationship $\Rightarrow$ slope of the relationship is different than 0.

Relevant Questions

1. Is $\beta_1 = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
    - We can address this using hypothesis testing.
    - $H_0 : \beta_1 = 0$ (nothing is going on): The explanatory variable is not a significant predictor of the response variable, i.e. no relationship $\Rightarrow$ slope of the relationship is 0.
    - $H_a : \beta_1 \neq 0$ (something going on): The explanatory variable is a significant predictor of the response variable, i.e. relationship $\Rightarrow$ slope of the relationship is different than 0.
2. What is the range of possible values that β_1 might take on?
    - We can use confidence interval for β_1 to answer this.

Relevant Questions

1. Is $\beta_1 = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
   - We can address this using hypothesis testing.
   - $H_0 : \beta_1 = 0$ (nothing is going on): The explanatory variable is not a significant predictor of the response variable, i.e. no relationship $\Rightarrow$ slope of the relationship is 0.
   - $H_a : \beta_1 \neq 0$ (something going on): The explanatory variable is a significant predictor of the response variable, i.e. relationship $\Rightarrow$ slope of the relationship is different than 0.
2. What is the range of possible values that β_1 might take on?
   - We can use confidence interval for β_1 to answer this.
3. What is the range of possible values of $E(Y)$ for given value X?
   - We can use confidence intervals for $E(Y)$.

Relevant Questions

1. Is $\beta_1 = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
    - We can address this using hypothesis testing.
    - $H_0 : \beta_1 = 0$ (nothing is going on): The explanatory variable is not a significant predictor of the response variable, i.e. no relationship $\Rightarrow$ slope of the relationship is 0.
    - $H_a : \beta_1 \neq 0$ (something going on): The explanatory variable is a significant predictor of the response variable, i.e. relationship $\Rightarrow$ slope of the relationship is different than 0.
2. What is the range of possible values that β_1 might take on?
    - We can use confidence interval for β_1 to answer this.
3. What is the range of possible values of $E(Y)$ for given value X?
    - We can use confidence intervals for $E(Y)$.
4. What is the range of possible values of Y for a given value of X?
    - We can use confidence intervals for Y.

1. Is $\beta_1 = 0$? i.e. is X an important variable?

- We use a hypothesis test to answer this question.
- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

1. Is $\beta_1 = 0$? i.e. is X an important variable?

- We use a hypothesis test to answer this question.
- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

use a t-statistic in inference for regression

b_1

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

SE_{b_1}

t-statistic for the slope: $T = \dfrac{b_1 - 0}{SE_{b_1}}$ $\qquad df = n - 2$

1. Is $\beta_1 = 0$? i.e. is X an important variable?

- We use a hypothesis test to answer this question.
- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

use a t-statistic in inference for regression

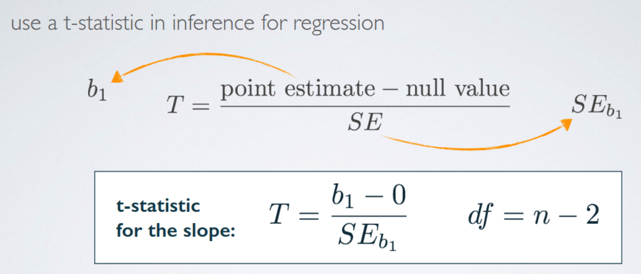
$$T = \frac{\overbrace{\text{point estimate}}^{b_1} - \text{null value}}{\underbrace{SE}_{SE_{b_1}}}$$

t-statistic for the slope: $T = \dfrac{b_1 - 0}{SE_{b_1}}$ $\qquad df = n - 2$

Remember: *We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.*

1. Is $\beta_1 = 0$? i.e. is X an important variable?

- We use a hypothesis test to answer this question.
- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.



Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, β_0 and β_1.

1. Is $\beta_1 = 0$? i.e. is X an important variable?

```
> summary(lm(price~weight))

Call:
lm(formula = price ~ weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-85.159 -21.448  -0.869  18.972  79.370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -259.63      17.32  -14.99   <2e-16 ***
weight       3721.02      81.79   45.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.84 on 46 degrees of freedom
Multiple R-squared:  0.9783,    Adjusted R-squared:  0.9778 
F-statistic:  2070 on 1 and 46 DF,  p-value: < 2.2e-16
```

2. Confidence Interval for β_1

- We not only care about whether $\beta_1 = 0$, but also what exact values β_1 takes.
- We can calculate confidence interval for β_1 using the same idea as for μ.

2. Confidence Interval for β_1

- We not only care about whether $\beta_1 = 0$, but also what exact values β_1 takes.
- We can calculate confidence interval for β_1 using the same idea as for μ.
- Here we see that, for every carat increase in weight, the ring expects to cost between \$3556.398 and \$3885.651 more.

```
> confint(lm(price~weight), "weight", level=0.95)
          2.5 %   97.5 %
weight 3556.398 3885.651
```

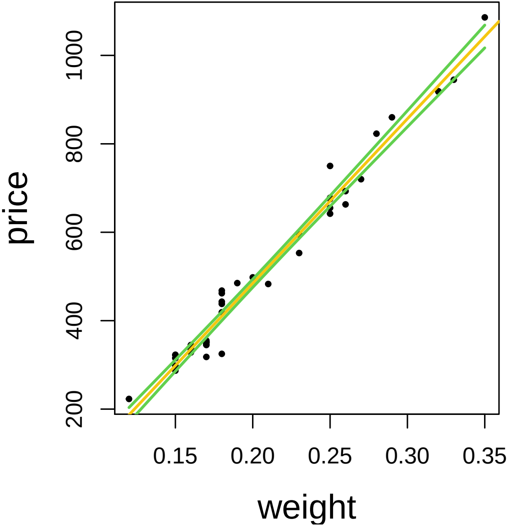
Testing via Confidence Interval

- Should the hypothesis $H_0 : \beta_1 = 0$ be rejected?
- Should the hypothesis $H_0 : \beta_1 = 3800$ be rejected?

```
> confint(lm(price~weight), "weight", level=0.95)
          2.5 %   97.5 %
weight 3556.398 3885.651
```

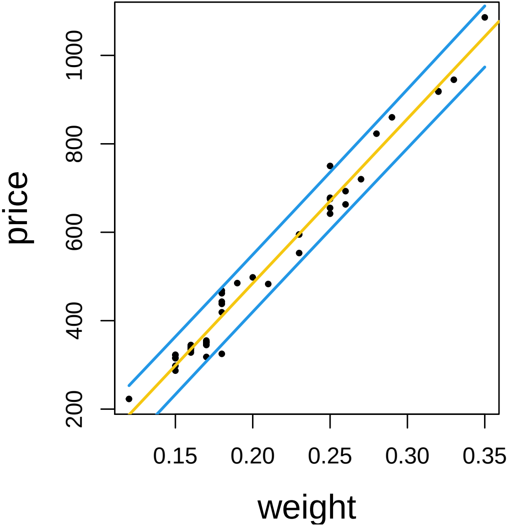
3. Confidence Interval for $E(y|x) = \beta_0 + \beta_1 x$

- What is the *average price for all* rings with 1/4 carat diamonds?

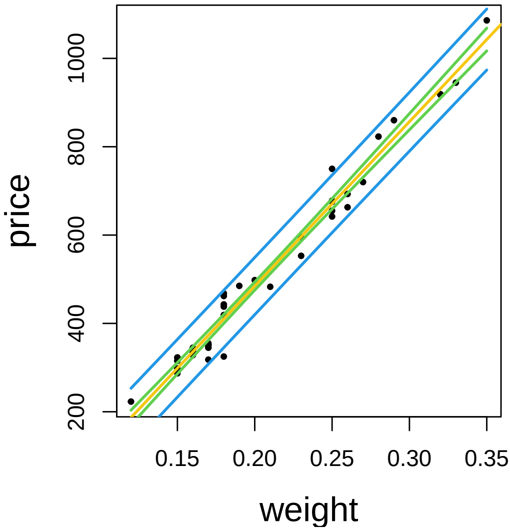


4. Prediction Interval for $y = \beta_0 + \beta_1 x + \epsilon$

- How much might pay for a *specific* ring with a 1/4 carat diamond?



Comparison



Interpretation of RMSE

- RMSE is especially important in regression.
- If the simple linear regression model holds, i.e.
  - Linear relationship,
  - Independence among the residuals,
  - The residuals have constant variance,
  - Normally distributed residuals,
- then we have the following approximations:
  - 68% of the observed y_i lies within $1 \times$RMSE of the fitted $\hat{y}_i$
  - 95% of the observed y_i lies within $2 \times$RMSE of the fitted $\hat{y}_i$
  - 99.7% of the observed y_i lies within $3 \times$RMSE of the fitted $\hat{y}_i$

Outline

1. Simple Linear Regression
   - Correlation
   - Least Squares Line
   - R^2
   - Model Diagnostics
   - Inference

2. Multiple Linear Regression
   - Model, Estimate, and Diagnostics
   - Inference
   - Collinearity
   - Categorical Explanatory Variables

Outline

Multiple Linear Regression

- Simple linear regression: Bivariate - *two variables*: y and x

Multiple Linear Regression

- Simple linear regression: Bivariate - *two variables*: y and x
- Multiple linear regression: *Multiple variables*: y and $x_1, x_2, \cdots$

Explaining and Predicting Fuel Efficiency

- The dataset contains characteristics of various makes and models of cars from the 2003 and 2004 model years.
- Variables in the data set include:
  - MPG City, Make/Model, Weight, Cargo, Seating, Horsepower, Price, ...

Explaining and Predicting Fuel Efficiency

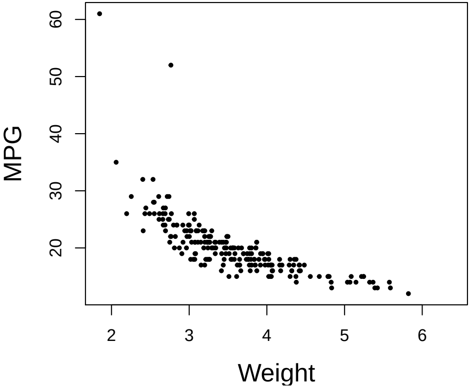
- The dataset contains characteristics of various makes and models of cars from the 2003 and 2004 model years.
- Variables in the data set include:
  - MPG City, Make/Model, Weight, Cargo, Seating, Horsepower, Price, ...
- *Questions of interest*:
  - What is the predicted mileage for a car with 4,000 lb. and 200 horsepower?
  - How much does my 200-pound brother owe me after riding with me for 3,000 miles?

Explaining and Predicting Fuel Efficiency

- The dataset contains characteristics of various makes and models of cars from the 2003 and 2004 model years.
- Variables in the data set include:
  - MPG City, Make/Model, Weight, Cargo, Seating, Horsepower, Price, ...
- *Questions of interest*:
  - What is the predicted mileage for a car with 4,000 lb. and 200 horsepower?
  - How much does my 200-pound brother owe me after riding with me for 3,000 miles?
- To get started, let's consider using simple regression to model the effect of *Weight* (measured in thousands of pounds) on *MPG* City (miles per gallon in urban driving).

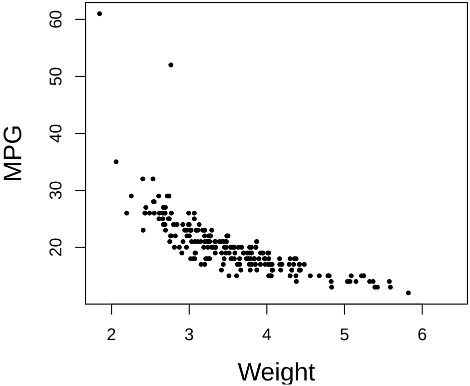
Regress MPG City on Weight

- We begin by examining a scatterplot of *MPG* City on *Weight*.



Regress MPG City on Weight

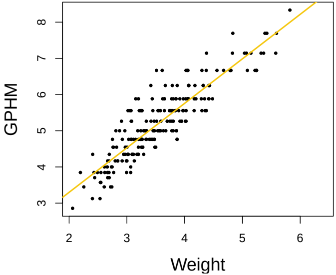
- We begin by examining a scatterplot of *MPG* City on *Weight*.



- Two obvious *outliers* – hybrid cars
- *Nonlinear* relationship

Regress GPHM on Weight

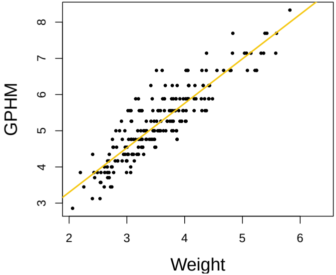
- We exclude the two outliers, and transform MPG City to $GPHM = 100/MPG$.




```
> summary(lm(GPHM~Weight,data = cars))

Call:
lm(formula = GPHM ~ Weight, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83482 -0.30899 -0.06211  0.23245  1.50875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.84114    0.15333   5.486 1.14e-07 ***
Weight       1.22915    0.04205  29.233  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4564 on 217 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7975,    Adjusted R-squared:  0.7966
F-statistic: 854.6 on 1 and 217 DF,  p-value: < 2.2e-16
```

Regress GPHM on Weight

- We exclude the two outliers, and transform MPG City to $GPHM = 100/MPG$.




```
> summary(lm(GPHM~Weight,data = cars))

Call:
lm(formula = GPHM ~ Weight, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83482 -0.30899 -0.06211  0.23245  1.50875 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.84114    0.15333   5.486 1.14e-07 ***
Weight       1.22915    0.04205  29.233  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4564 on 217 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7975,	Adjusted R-squared:  0.7966 
F-statistic: 854.6 on 1 and 217 DF,  p-value: < 2.2e-16
```

- For 1000 pound increase in weight, gas consumption increases *1.23 gallons* every 100 miles.

Regress GPHM on more variables

- Regress GPHM on Weight for all the cars.

Regress GPHM on more variables

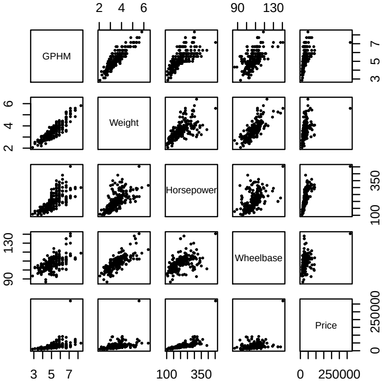
- Regress GPHM on Weight for all the cars.
- However, is 4,000 pounds the only difference between the Corolla and the Rolls?

Regress GPHM on more variables

- Regress GPHM on Weight for all the cars.
- However, is 4,000 pounds the only difference between the Corolla and the Rolls?
- *Other factors contribute* to GPHM as well.
  - Examine GPHM, Weight, Horsepower, Wheelbase, and Price.

More Factors

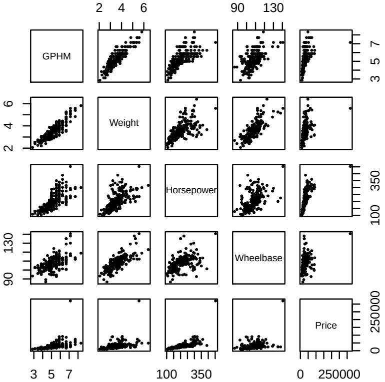
- Scatterplot matrix and correlation matrix for the five variables



	GPHM	Weight	Horsepower	Wheelbase	Price
GPHM	1.0000000	0.8930267	0.7229165	0.5705430	0.4541867
Weight	0.8930267	1.0000000	0.6355802	0.7637742	0.4356994
Horsepower	0.7229165	0.6355802	1.0000000	0.4749461	0.6858165
Wheelbase	0.5705430	0.7637742	0.4749461	1.0000000	0.3652725
Price	0.4541867	0.4356994	0.6858165	0.3652725	1.0000000

More Factors

- Scatterplot matrix and correlation matrix for the five variables



	GPHM	Weight	Horsepower	Wheelbase	Price
GPHM	1.0000000	0.8930267	0.7229165	0.5705430	0.4541867
Weight	0.8930267	1.0000000	0.6355802	0.7637742	0.4356994
Horsepower	0.7229165	0.6355802	1.0000000	0.4749461	0.6858165
Wheelbase	0.5705430	0.7637742	0.4749461	1.0000000	0.3652725
Price	0.4541867	0.4356994	0.6858165	0.3652725	1.0000000

- These results summarize the pairwise associations between the five variables. The Rolls is a big outlier. After weight, *horsepower* is most strongly associated with GPHM.

Multivariate Regression Model

- Consider the *joint* effect of Weight and Horsepower on GPHM

```
> summary(lm(GPHM~Weight+Horsepower,data = cars))

Call:
lm(formula = GPHM ~ Weight + Horsepower, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.13561 -0.26998 -0.04683  0.23506  1.61326 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.8085502  0.1375390   5.879 1.55e-08 ***
Weight      1.0011798  0.0488281  20.504  < 2e-16 ***
Horsepower  0.0041641  0.0005669   7.346 4.14e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4092 on 216 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.838,	Adjusted R-squared:  0.8365 
F-statistic: 558.6 on 2 and 216 DF,  p-value: < 2.2e-16
```

Single Factor vs Two Factors

```
> summary(lm(GPHM~Weight,data = cars))

Call:
lm(formula = GPHM ~ Weight, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83482 -0.30899 -0.06211  0.23245  1.50875 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.84114    0.15333   5.486 1.14e-07 ***
Weight       1.22915    0.04205  29.233  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4564 on 217 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7975,	Adjusted R-squared:  0.7966 
F-statistic: 854.6 on 1 and 217 DF,  p-value: < 2.2e-16
```

```
> summary(lm(GPHM~Weight+Horsepower,data = cars))

Call:
lm(formula = GPHM ~ Weight + Horsepower, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.13561 -0.26998 -0.04683  0.23506  1.61326 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.8085502  0.1375390   5.879 1.55e-08 ***
Weight      1.0011798  0.0488281  20.504  < 2e-16 ***
Horsepower  0.0041641  0.0005669   7.346 4.14e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4092 on 216 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.838,	Adjusted R-squared:  0.8365 
F-statistic: 558.6 on 2 and 216 DF,  p-value: < 2.2e-16
```

Partial versus Marginal Regression Coefficients

- The LS regression line is

$$Est\ \ GPHM \ = \ 0.809 + 1.001\,Weight + 0.00416\,Horsepower$$

- The coefficient $b_1 = 1.001$ estimates the average GPHM increase per thousand pound increase in Weight *for a fixed level of horsepower*.
- This interpretation implies that we are comparing the fuel consumption of cars of different weights, but *identical horsepower*.

Partial versus Marginal Regression Coefficients

- The LS regression line is

$$\textit{Est GPHM} = 0.809 + 1.001\,\textit{Weight} + 0.00416\,\textit{Horsepower}$$

- The coefficient $b_1 = 1.001$ estimates the average GPHM increase per thousand pound increase in Weight *for a fixed level of horsepower*.
- This interpretation implies that we are comparing the fuel consumption of cars of different weights, but *identical horsepower*.
- With other explanatory variables in the equation, b_1 is called a partial regression coefficient.

Partial versus Marginal Regression Coefficients

- The simple linear regression with only Weight is

$$\textit{Estimated GPHM} = 0.841 + 1.229\textit{Weight}$$

- The interpretation of $b_1 = 1.229$ in this simple regression is an increase due to Weight that *averages over all other differences including horsepower*.

Partial versus Marginal Regression Coefficients

- The simple linear regression with only Weight is

$$\textit{Estimated GPHM} = 0.841 + 1.229\,\textit{Weight}$$

- The interpretation of $b_1 = 1.229$ in this simple regression is an increase due to Weight that *averages over all other differences including horsepower*.
- With no other explanatory variables in the equation, b_1 here is called a marginal regression coefficient.

Partial versus Marginal Regression Coefficients

- The simple linear regression with only Weight is

$$Estimated\ GPHM = 0.841 + 1.229 Weight$$

- The interpretation of $b_1 = 1.229$ in this simple regression is an increase due to Weight that *averages over all other differences including horsepower*.
- With no other explanatory variables in the equation, b_1 here is called a marginal regression coefficient.
- *Partial* regression coefficients adjust for the effects of other variables, whereas *marginal* regression coefficients average over the effects of other variables.

Least-Squares Estimation and Prediction

- Consider

$$\hat{y}_i = b_0 + b_1 x_{i1} + \dots + b_K x_{iK}$$

- Find $b_0, b_1, \dots, b_K$ to minimize

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \dots - b_K x_{iK})^2$$

Least-Squares Estimation and Prediction

- Consider

$$\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_K x_{iK}$$

- Find $b_0, b_1, ..., b_K$ to minimize

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - ... - b_K x_{iK})^2$$

- Matrix form
  - Y: response vector
  - X: design matrix with variables as columns
  - B: coefficient vector

Least-Squares Estimation and Prediction

- Consider

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_K x_{iK}$$

- Find $b_0, b_1, \ldots, b_K$ to minimize

$$SSE = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_K x_{iK})^2$$

- Matrix form
  - Y: response vector
  - X: design matrix with variables as columns
  - B: coefficient vector
- Normal equation

$$B = (X^T X)^{-1} X^T Y$$

- Prediction

$$\hat{Y} = X(X^T X)^{-1} X^T Y$$

Multiple Linear Regression Model

- In our multiple regression model, the data $(x_{11}, ..., x_{1K}, y_1), ..., (x_{n1}, ..., x_{nK}, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + \epsilon_i$$

where ϵ_i iid $\sim N(0, \sigma^2)$

Multiple Linear Regression Model

- In our multiple regression model, the data $(x_{11}, \dots, x_{1K}, y_1), \dots, (x_{n1}, \dots, x_{nK}, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_K x_{iK} + \epsilon_i$$

  where ϵ_i iid $\sim N(0, \sigma^2)$
- Least squares regression line:
  - Fitted values: $\hat{y}_i = b_0 + b_1 x_{i1} + \dots + b_K x_{iK}$
  - Residuals: $e_i = y_i - \hat{y}_i$

Multiple Linear Regression Model

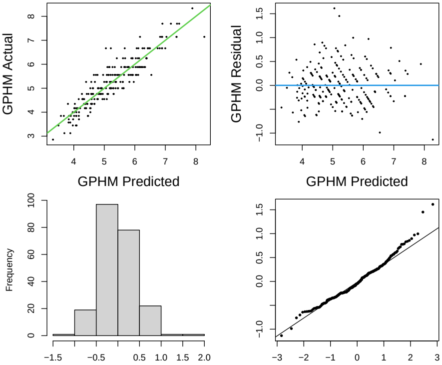
- In our multiple regression model, the data $(x_{11}, ..., x_{1K}, y_1), ..., (x_{n1}, ..., x_{nK}, y_n)$ are a realization of

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + \epsilon_i$$

  where ϵ_i iid $\sim N(0, \sigma^2)$
- Least squares regression line:
  - Fitted values: $\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_K x_{iK}$
  - Residuals: $e_i = y_i - \hat{y}_i$
- *RMSE* can be used to estimate σ

Residual Diagnostics

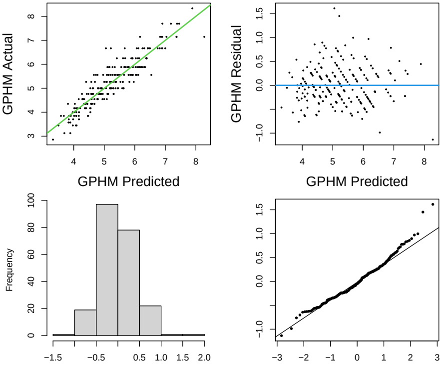
- For the multiple regression of GPHM on Weight and Horsepower,



- We want to see if there is *systematic pattern* in the residuals.

Residual Diagnostics

- For the multiple regression of GPHM on Weight and Horsepower,



- We want to see if there is *systematic pattern* in the residuals.
- R^2 equals to the squared *correlation between y and $\hat{y}$*.

Outline

Inferences in Multiple Linear Regression

- Once we have checked the *assumptions* of the multiple linear regression model, we can proceed to inference.
- Tests and confidence intervals used in simple linear regression *generalize naturally* to multiple linear regression.
- Inferences for the multiple linear regression refer to *partial* slopes rather than marginal slopes.

Relevant Questions

1. Is $\beta_k = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
   - If we can't be sure that $\beta_k \neq 0$, then no point in using X_k as one predictor.
   - We can address this using hypothesis testing.
2. What is the range of possible values that β_k might take on?
   - We can use confidence interval for β_k to answer this.

Relevant Questions

1. Is $\beta_k = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
   - If we can't be sure that $\beta_k \neq 0$, then no point in using X_k as one predictor.
   - We can address this using hypothesis testing.
2. What is the range of possible values that β_k might take on?
   - We can use confidence interval for β_k to answer this.
3. What is the range of possible values of $E(Y)$ for given value X?
   - We can use confidence intervals for $E(Y)$.
4. What is the range of possible values of Y for a given value of X?
   - We can use confidence intervals for Y.

Relevant Questions

1. Is $\beta_k = 0$, i.e. is the explanatory variable a significant predictor of the response variable?
   - If we can't be sure that $\beta_k \neq 0$, then no point in using X_k as one predictor.
   - We can address this using hypothesis testing.
2. What is the range of possible values that β_k might take on?
   - We can use confidence interval for β_k to answer this.
3. What is the range of possible values of $E(Y)$ for given value X?
   - We can use confidence intervals for $E(Y)$.
4. What is the range of possible values of Y for a given value of X?
   - We can use confidence intervals for Y.
5. *Is any predictor useful in predicting Y? i.e. $\beta_k = 0$ for all k?*
   - We can use F test.

Cars Example

- Regress GPHM on Weight, Horsepower, Wheelbase and Price

```
> summary(lm(GPHM~Weight+Horsepower+Wheelbase+Price,data = cars))

Call:
lm(formula = GPHM ~ Weight + Horsepower + Wheelbase + Price,
    data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86423 -0.24880 -0.01287  0.19905  1.74530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.326e+00  4.129e-01   8.057 5.44e-14 ***
Weight       1.274e+00  6.043e-02  21.080  < 2e-16 ***
Horsepower   4.740e-03  6.366e-04   7.446 2.32e-12 ***
Wheelbase   -3.279e-02  4.950e-03  -6.623 2.79e-10 ***
Price       -2.372e-06  1.373e-06  -1.728   0.0855 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3708 on 214 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8682,	Adjusted R-squared:  0.8658
F-statistic: 352.5 on 4 and 214 DF,  p-value: < 2.2e-16
```

- The t ratio for β_k measures the effect of adding x_k last.
- What would you conclude about the effect of either addition?

Cars Example

- Regress GPHM on Weight, Horsepower, Wheelbase and Price

```
> summary(lm(GPHM~Weight+Horsepower+Wheelbase+Price,data = cars))

Call:
lm(formula = GPHM ~ Weight + Horsepower + Wheelbase + Price,
    data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86423 -0.24880 -0.01287  0.19905  1.74530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.326e+00  4.129e-01   8.057 5.44e-14 ***
Weight       1.274e+00  6.043e-02  21.080  < 2e-16 ***
Horsepower   4.740e-03  6.366e-04   7.446 2.32e-12 ***
Wheelbase   -3.279e-02  4.950e-03  -6.623 2.79e-10 ***
Price       -2.372e-06  1.373e-06  -1.728   0.0855 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3708 on 214 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8682,    Adjusted R-squared:  0.8658
F-statistic: 352.5 on 4 and 214 DF,  p-value: < 2.2e-16
```

- The t ratio for β_k measures the effect of adding x_k last.
- What would you conclude about the effect of either addition?
- If Price is removed, the other t ratios and p-values will change. The regression must then be rerun to get a new set of t ratios.

Cars Example

- Regress GPHM on Weight, Horsepower, Wheelbase and Price

```
> summary(lm(GPHM~Weight+Horsepower+Wheelbase+Price,data = cars))

Call:
lm(formula = GPHM ~ Weight + Horsepower + Wheelbase + Price,
    data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86423 -0.24880 -0.01287  0.19905  1.74530

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.326e+00  4.129e-01   8.057 5.44e-14 ***
Weight       1.274e+00  6.043e-02  21.080  < 2e-16 ***
Horsepower   4.740e-03  6.366e-04   7.446 2.32e-12 ***
Wheelbase   -3.279e-02  4.950e-03  -6.623 2.79e-10 ***
Price       -2.372e-06  1.373e-06  -1.728   0.0855 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3708 on 214 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8682,    Adjusted R-squared:  0.8658
F-statistic: 352.5 on 4 and 214 DF,  p-value: < 2.2e-16
```

- The t ratio for β_k measures the effect of adding x_k last.
- What would you conclude about the effect of either addition?
- If Price is removed, the other t ratios and p-values will change. The regression must then be rerun to get a new set of t ratios.
- The increase in R^2 due to adding x_k last is said to be significant when the t ratio for β_k is significant.

5. Is At Least One $\beta_k \neq 0$?

- Regress GPHM on Weight, Horsepower, Wheelbase, and Price

```
F-statistic: 352.5 on 4 and 214 DF,  p-value: < 2.2e-16

> anova(lm(GPHM~Weight+Horsepower+Wheelbase+Price,data = cars))
Analysis of Variance Table

Response: GPHM
           Df  Sum Sq Mean Sq   F value    Pr(>F)    
Weight      1 178.044 178.044 1295.1951 < 2.2e-16 ***
Horsepower  1   9.037   9.037   65.7390 3.945e-14 ***
Wheelbase   1   6.345   6.345   46.1563 1.064e-10 ***
Price       1   0.410   0.410    2.9856   0.08545 .  
Residuals 214  29.417   0.137                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $H_0 : \beta_k = 0$ for all k.

5. Is At Least One $\beta_k \neq 0$?

- Regress GPHM on Weight, Horsepower, Wheelbase, and Price

```
F-statistic: 352.5 on 4 and 214 DF,  p-value: < 2.2e-16

> anova(lm(GPHM~Weight+Horsepower+Wheelbase+Price,data = cars))
Analysis of Variance Table

Response: GPHM
            Df  Sum Sq Mean Sq   F value    Pr(>F)    
Weight       1 178.044 178.044 1295.1951 < 2.2e-16 ***
Horsepower   1   9.037   9.037   65.7390 3.945e-14 ***
Wheelbase    1   6.345   6.345   46.1563 1.064e-10 ***
Price        1   0.410   0.410    2.9856   0.08545 .  
Residuals  214  29.417   0.137                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $H_0 : \beta_k = 0$ for all k.

$$F = \frac{(TSS - SSE)/K}{SSE/(n-K-1)} = \frac{(178.04 + 9.04 + 6.35 + 0.41)/4}{29.42/214} = 352.5 \sim F_{K,n-K-1}$$

5. Is At Least One $\beta_k \neq 0$?

- Regress GPHM on Weight, Horsepower, Wheelbase, and Price

```
F-statistic: 352.5 on 4 and 214 DF,  p-value: < 2.2e-16

> anova(lm(GPHM~Weight+Horsepower+Wheelbase+Price,data = cars))
Analysis of Variance Table

Response: GPHM
           Df  Sum Sq Mean Sq   F value    Pr(>F)    
Weight      1 178.044 178.044 1295.1951 < 2.2e-16 ***
Horsepower  1   9.037   9.037   65.7390 3.945e-14 ***
Wheelbase   1   6.345   6.345   46.1563 1.064e-10 ***
Price       1   0.410   0.410    2.9856   0.08545 .  
Residuals 214  29.417   0.137                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $H_0 : \beta_k = 0$ for all k.

$$F = \frac{(TSS - SSE)/K}{SSE/(n-K-1)} = \frac{(178.04 + 9.04 + 6.35 + 0.41)/4}{29.42/214} = 352.5 \sim F_{K,n-K-1}$$

- The ANOVA table supplies a highly significant F-ratio.
  - This model explains statistically significant variation in Y.
  - At least one β_k is not zero, i.e. at least one of the X_k's is useful in predicting Y.

Multiple Linear Regression Review

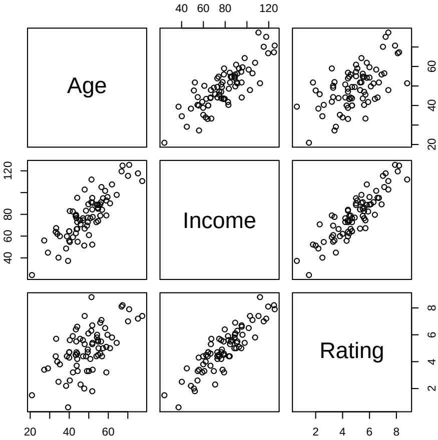
- The multiple linear regression extends the simple linear regression, allowing for *more predictors*.
- Under the model assumptions, we can again use standard errors to form confidence intervals and test hypotheses.
- To assess the model assumptions, new *diagnostic plots* include plot of fitted value on actual values of the response, and plot of residuals on fitted values.
- The *ANOVA Table* allows you to look at the importance of several factors simultaneously.

Outline

Example: Market Segmentation

- A *marketing* project identified a list of affluent customers for a new phone.
- Should the company target promotion towards the *younger or older* members of this list?
- To answer this question, the marketing firm obtained a sample of 75 consumers and asked them to rate their *"likelihood of purchase"* on a scale of 1 to 10.
- *Age and Income* of consumers were also recorded.

Correlation Among Variables



Correlation

	Age	Income	Rating
Age	1.000	0.828	0.586
Income	0.828	1.000	0.884
Rating	0.586	0.884	1.000

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

- *MRM of Rating, on both variables*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.512374	0.355004	1.443	0.153
Age	-0.071448	0.012576	-5.682	2.65e-07
Income	*0.100591*	0.006491	15.498	$< 2e-16$

Smartphone

- *SRM of Rating, one variable at a time*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.49004	0.73414	0.668	0.507
Age	*0.09002*	0.01456	6.181	3.3e-08

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.598441	0.354155	-1.69	0.0953
Income	*0.070039*	0.004344	16.12	$< 2e-16$

- *MRM of Rating, on both variables*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.512374	0.355004	1.443	0.153
Age	-0.071448	0.012576	-5.682	2.65e-07
Income	*0.100591*	0.006491	15.498	$< 2e-16$

- We need to understand why the slope of *Age* is positive in the simple regression but negative in the multiple regression.
- Given the context, the positive marginal slope is probably more surprising than the negative partial slope.

Collinearity: Highly Correlated X Variables

- MRM allows the use of correlated explanatory variables.
- *Collinearity* occurs when the correlations among the X variables are large.

Collinearity: Highly Correlated X Variables

- MRM allows the use of correlated explanatory variables.
- *Collinearity* occurs when the correlations among the X variables are large.
- As the correlation among these variables grows, it becomes difficult for regression to separate the partial effects of different variables.
  - Highly correlated X variables tend to change together, making it *difficult to estimate* the partial slope.
  - *Difficulties interpreting* the model

Customer Segmentation

- The figure shows regression lines fit within three subsets:

low incomes ($< \$45$K)

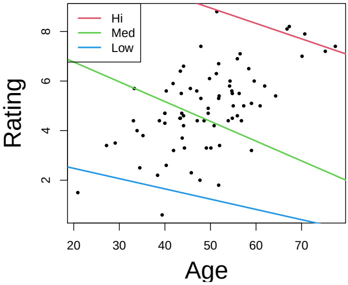
	Estimate	Std. Error	t value	$Pr(>|t|)$
(Intercept)	3.30845	3.42190	0.967	0.436
Age	-0.04144	0.10786	-0.384	0.738

moderate incomes ($\$70$K $\sim$ $\$80$K)

	Estimate	Std. Error	t value	$Pr(>|t|)$
(Intercept)	8.36412	2.34772	3.563	0.0026
Age	-0.07978	0.04791	-1.665	0.1153

high incomes ($> \$110$K)

	Estimate	Std. Error	t value	$Pr(>|t|)$
(Intercept)	12.07081	1.28999	9.357	0.000235
Age	-0.06243	0.01873	-3.332	0.020727



- The simple regression slopes are negative in each case, as in the *multiple linear regression*.

Customer Segmentation

- The figure shows regression lines fit within three subsets:

low incomes ($< \$45K$)

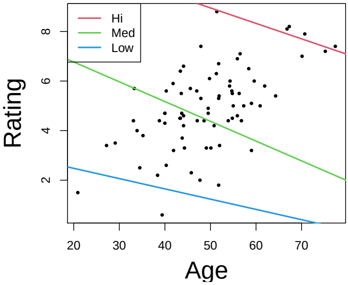
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	3.30845	3.42190	0.967	0.436
Age	-0.04144	0.10786	-0.384	0.738

moderate incomes ($\$70K \sim \$80K$)

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	8.36412	2.34772	3.563	0.0026
Age	-0.07978	0.04791	-1.665	0.1153

high incomes ($> \$110K$)

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	12.07081	1.28999	9.357	0.000235
Age	-0.06243	0.01873	-3.332	0.020727

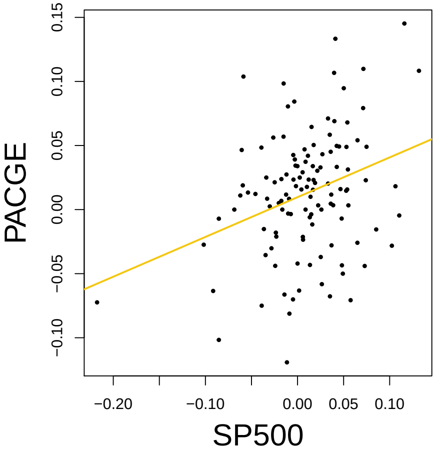


- The simple regression slopes are negative in each case, as in the *multiple linear regression*.
- Based on these results, how should the marketing firm direct their promotional efforts?

The Market Model

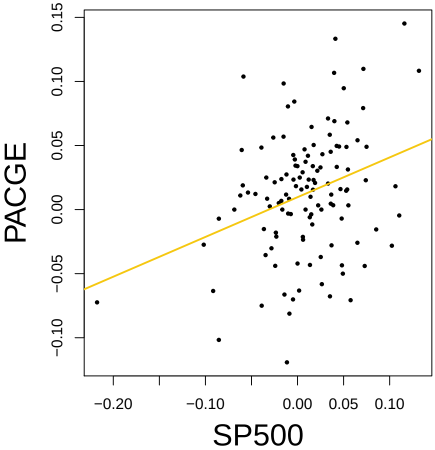
- We consider simple linear regression of
  - exPACGE on exSP500, the excess returns of PACGE and SP500 over TBill30
  - exPACGE on exVW, the excess returns of PACGE and VW over TBill30
- Also, consider multiple linear regression of exPACGE on both exSP500 and exVW

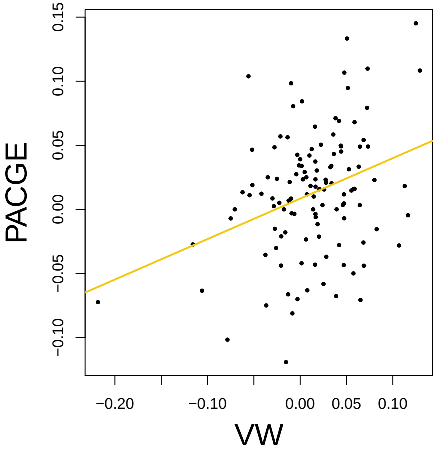
SRM of exPACGE on either the exSP500 or exVW



	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

SRM of exPACGE on either the exSP500 or exVW





	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

- Very similar results.

Regress exPACGE on both exSP500 and exVW

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

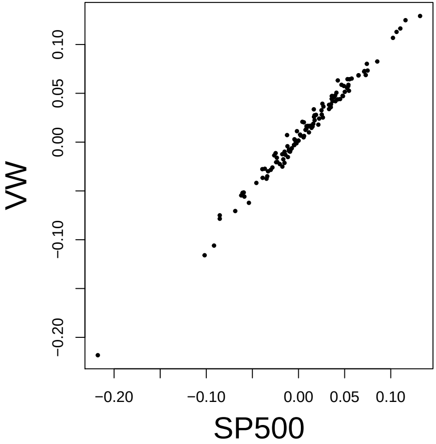
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.005448	0.005119	1.064	0.289
SP500	-0.821098	0.749946	-1.095	0.276
VW	1.111498	0.731784	1.519	0.132
ANOVA	F-statistic	7.513	p-value	0.0008547

Regress exPACGE on both exSP500 and exVW

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.005448	0.005119	1.064	0.289
SP500	-0.821098	0.749946	-1.095	0.276
VW	1.111498	0.731784	1.519	0.132
ANOVA	F-statistic	7.513	p-value	0.0008547

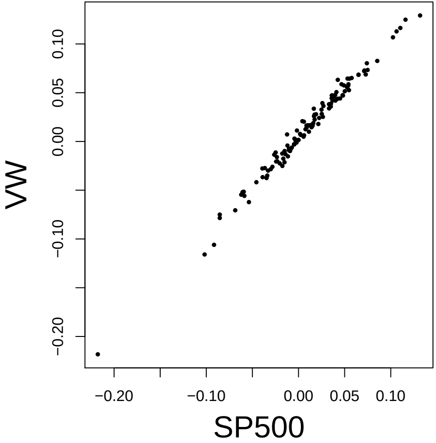


Regress exPACGE on both exSP500 and exVW

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.009682	0.004317	2.243	0.026803
SP500	*0.310295*	0.087490	3.547	*0.000562*
ANOVA	F-statistic	12.58	p-value	*0.0005623*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.008371	0.004371	1.915	0.057918
VW	*0.315696*	0.084970	3.715	*0.000313*
ANOVA	F-statistic	13.8	p-value	*0.0003126*

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.005448	0.005119	1.064	0.289
SP500	-0.821098	0.749946	-1.095	0.276
VW	1.111498	0.731784	1.519	0.132
ANOVA	F-statistic	7.513	p-value	0.0008547



- Huge Collinearity!!!

The F Test and Correlated Predictors

- *Seemingly contradiction* between
  - Overall F Ratio in the ANOVA Table
  - Individual p-value (T test) for each regression coefficient

The F Test and Correlated Predictors

- *Seemingly contradiction* between
  - Overall F Ratio in the ANOVA Table
  - Individual p-value (T test) for each regression coefficient
- The overall F Ratio comes in handy when the explanatory variables in a regression are correlated.
  - *Overall F Ratio*: whether at least one of the X variables is significant, leaving out the other ones
  - *Individual T test*: whether each individual X variable is significant, having included the other ones
- When the predictors are highly correlated (i.e. *high collinearity*), they may contradict each other.

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is *R^2 from regressing x_k on the other x's.*
- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated,

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is R^2 *from regressing* x_k *on the other* x*'s.*

- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated, VIF $= 1$.
- If the x's are correlated,

Measuring Collinearity: Variance Inflation Factor (VIF)

- The *VIF* is defined as

$$VIF(b_k) = \frac{1}{1 - R_k^2}$$

  where R_k^2 is *R^2 from regressing x_k on the other x's.*
- The VIF is the *ratio* of the variation that was originally in each explanatory variable to the variation that remains after removing the effects of the other explanatory variables.
- If the x's are uncorrelated, VIF $= 1$.
- If the x's are correlated, VIF can be much larger than 1.

VIF Results

- For market example

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.005448	0.005119	1.064	0.289	
SP500	-0.821098	0.749946	-1.095	*0.276*	74.29672
VW	1.111498	0.731784	1.519	*0.132*	74.29672

- For Customer Segmentation

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.512374	0.355004	1.443	0.153	
Age	-0.071448	0.012576	-5.682	*2.65e-07*	3.188591
Income	0.100591	0.006491	15.498	$< 2e-16$	3.188591

VIF Results

- For market example

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.005448	0.005119	1.064	0.289	
SP500	-0.821098	0.749946	-1.095	*0.276*	74.29672
VW	1.111498	0.731784	1.519	*0.132*	74.29672

- For Customer Segmentation

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	VIF
(Intercept)	0.512374	0.355004	1.443	0.153	
Age	-0.071448	0.012576	-5.682	*2.65e-07*	3.188591
Income	0.100591	0.006491	15.498	$< 2e - 16$	3.188591

- The VIF answers a very handy question when an explanatory variable is not statistically significant:
  - Is this explanatory variable simply not useful, or is it just redundant?

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's
- The *F Ratio* detects statistical significance that can be disguised by collinearity.
  - The F ratio allows you to look at the importance of several factors simultaneously.
  - When predictors are collinear, the F test reveals their net effect, rather than trying to separate their effects as a t ratio does.

Summary: Collinearity

- *Collinearity* is the presence of "substantial" correlation among the explanatory variables (the X's) in a multiple regression.
  - Potential redundancy among the X's
- The *F Ratio* detects statistical significance that can be disguised by collinearity.
  - The F ratio allows you to look at the importance of several factors simultaneously.
  - When predictors are collinear, the F test reveals their net effect, rather than trying to separate their effects as a t ratio does.
- *VIF measures* the impact of collinearity on the coefficients of specific explanatory variables.

Summary: Collinearity

- Collinearity does *not violate* any assumption of the MRM, but it does make regression harder to interpret.
  - In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.

Summary: Collinearity

- Collinearity does *not violate* any assumption of the MRM, but it does make regression harder to interpret.
  - In the presence of collinearity, slopes become less precise and the effect of one predictor depends on the others that happen to be in the model.
- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. *parsimonious* model.

R^2 vs. adjusted R^2

- When any variable is added to the model, R^2 *increases*.

R^2 vs. adjusted R^2

- When <u>any</u> variable is added to the model, R^2 *increases*.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adj R^2 does not increase.

R^2 vs. adjusted R^2

- When any variable is added to the model, R^2 *increases*.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adj R^2 does not increase.
- R^2

$$R^2 = 1 - \frac{SSE}{TSS}$$

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{TSS/(n-1)}$$

where n is the number of cases and K is the number of predictors

R^2 vs. adjusted R^2

- When any variable is added to the model, *R^2 increases*.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adj R^2 does not increase.
- R^2

$$R^2 = 1 - \frac{SSE}{TSS}$$

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{TSS/(n-1)}$$

  where n is the number of cases and K is the number of predictors
- Because K is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors
- Therefore, we can choose models with higher R^2_{adj} over others.

R^2 vs. adjusted R^2

```
> summary(lm(PACGE~SP500+VW,data = stock))

Call:
lm(formula = PACGE ~ SP500 + VW, data = stock)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.117084 -0.025683  0.001373  0.029422  0.112175 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.005448   0.005119   1.064    0.289
SP500       -0.821098   0.749946  -1.095    0.276
VW           1.111498   0.731784   1.519    0.132

Residual standard error: 0.04591 on 116 degrees of freedom
Multiple R-squared:  0.1147,	Adjusted R-squared:  0.09942 
F-statistic: 7.513 on 2 and 116 DF,  p-value: 0.0008547

> 
> x3 <- rnorm(length(SP500))
> summary(lm(PACGE~SP500+VW+x3,data = stock))

Call:
lm(formula = PACGE ~ SP500 + VW + x3, data = stock)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.117041 -0.023896  0.004667  0.030164  0.108113 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006487   0.005151   1.259    0.210
SP500       -0.711646   0.750875  -0.948    0.345
VW           0.988744   0.733957   1.347    0.181
x3          -0.005476   0.003898  -1.405    0.163

Residual standard error: 0.04571 on 115 degrees of freedom
Multiple R-squared:  0.1296,	Adjusted R-squared:  0.1069 
F-statistic: 5.708 on 3 and 115 DF,  p-value: 0.001117
```

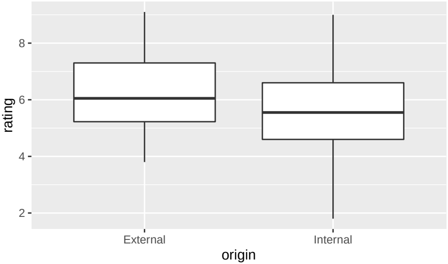
Outline

Example: Employee Performance Study

- "Which of two prospective job candidates should we hire for a position that pays \$80,000: the internal manager or the externally recruited manager?"
- Data set:
  - 150 managers: 88 internal and 62 external
  - *Manager Rating* is an evaluation score of the employee in their current job, indicating the "value" of the employee to the firm.
  - *Origin* is a categorical variable that identifies the managers as either External or Internal to indicate from where they were hired.
  - *Salary* is the starting salary of the employee when they were hired. It indicates what sort of job the person was initially hired to do. In the context of this example, it does not measure how well they did that job. That's measured by the rating variable.

Two-Sample Comparison: Manager Rating vs Origin

- *Origin*: a categorical variable.




```
        Welch Two Sample t-test

data:  rating by origin
t = 3.0484, df = 140.49, p-value = 0.00275
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 0.2517995 1.1810451
sample estimates:
mean in group External mean in group Internal
              6.320968               5.604545
```

- We can recognize a significant difference between the means via two-sample t-test.

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$\begin{aligned} y_{i|x=External} &= \mu_{External} + \epsilon_i \\ y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i \end{aligned}$$

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$\begin{aligned} y_{i|x=External} &= \mu_{External} + \epsilon_i \\ y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i \end{aligned}$$

- *In regression*
  - 'External' as the base
  - x_1 be the indicator function of being 'Internal', $I(Origin = Internal)$
  - $\beta_0 = \mu_{External}$
  - $\beta_1 = \mu_{Internal} - \mu_{External}$

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$\begin{aligned} y_{i|x=External} &= \mu_{External} + \epsilon_i \\ y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i \end{aligned}$$

- *In regression*
  - 'External' as the base
  - x_1 be the indicator function of being 'Internal', $I(Origin = Internal)$
  - $\beta_0 = \mu_{External}$
  - $\beta_1 = \mu_{Internal} - \mu_{External}$
- ANOVA model is the same as

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

One-way ANOVA

- Definition: regression model with one categorical variable.
- *ANOVA Model*

$$
\begin{aligned}
y_{i|x=External} &= \mu_{External} + \epsilon_i \\
y_{i|x=Internal} &= \mu_{Internal} + \epsilon_i
\end{aligned}
$$

- *In regression*
  - 'External' as the base
  - x_1 be the indicator function of being 'Internal', $I(Origin = Internal)$
  - $\beta_0 = \mu_{External}$
  - $\beta_1 = \mu_{Internal} - \mu_{External}$
- ANOVA model is the same as

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

- These two tests are *equivalent*

$$H_0 : \mu_{Internal} = \mu_{External} \text{ and } H_0 : \beta_1 = 0$$

Regress Manager Rating on Origin

```
Call:
lm(formula = rating ~ origin)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8045 -1.0169 -0.1045  0.9790  3.3955

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.3210     0.1839  34.372  < 2e-16 ***
originInternal  -0.7164     0.2401  -2.984  0.00333 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.448 on 148 degrees of freedom
Multiple R-squared:  0.05675,   Adjusted R-squared:  0.05037
F-statistic: 8.904 on 1 and 148 DF,  p-value: 0.00333
```

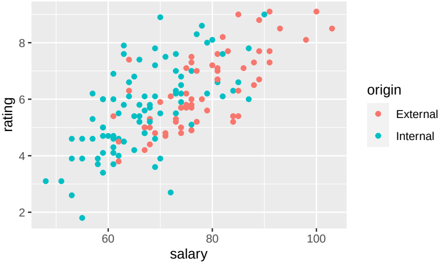
- The difference in the rating (-0.72) between internal and external managers is significant since the p-value $= .003 < .05$.
- In terms of regression, *Origin* explains significant variation in *Manager Rating*.

Regress Manager Rating on Origin

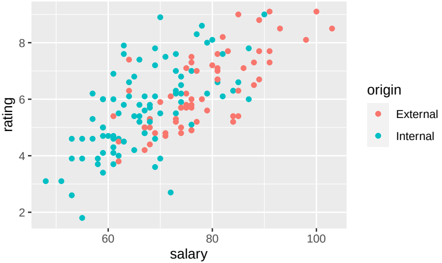
```
Call:
lm(formula = rating ~ origin)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8045 -1.0169 -0.1045  0.9790  3.3955 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      6.3210     0.1839  34.372  < 2e-16 ***
originInternal  -0.7164     0.2401  -2.984  0.00333 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.448 on 148 degrees of freedom
Multiple R-squared:  0.05675,	Adjusted R-squared:  0.05037 
F-statistic: 8.904 on 1 and 148 DF,  p-value: 0.00333
```

- The difference in the rating (-0.72) between internal and external managers is significant since the p-value $= .003 < .05$.
- In terms of regression, *Origin* explains significant variation in *Manager Rating*.
- Before we claim that the external candidate should be hired, is there a possible confounding variable, another explanation for the difference in rating?
- Let's explore the relationship between *Manager Rating and Salary*.

Scatterplot of Manager Rating vs. Salary

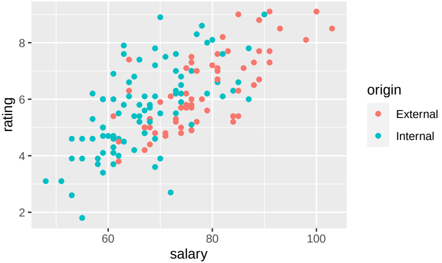


Scatterplot of Manager Rating vs. Salary



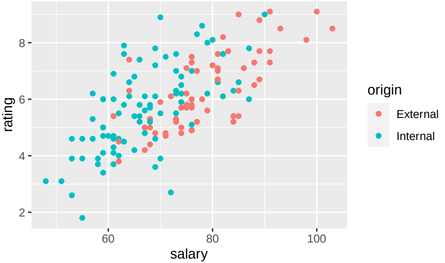
- (a) Salary is correlated with Manager Rating, and (b) that external managers were hired at higher salaries

Scatterplot of Manager Rating vs. Salary



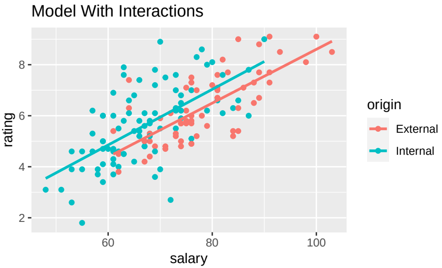
- (a) Salary is correlated with Manager Rating, and (b) that external managers were hired at higher salaries
- This combination indicates *confounding*: not only are we comparing internal vs. external managers; we are comparing internal managers hired into lower salary jobs with external managers placed into higher salary jobs.

Scatterplot of Manager Rating vs. Salary



- (a) Salary is correlated with Manager Rating, and (b) that external managers were hired at higher salaries
- This combination indicates *confounding*: not only are we comparing internal vs. external managers; we are comparing internal managers hired into lower salary jobs with external managers placed into higher salary jobs.
- *Easy fix*: compare only those whose starting salary near $80K. But that leaves too few data points for a reasonable comparison.

Separate Regressions of Manager Rating on Salary



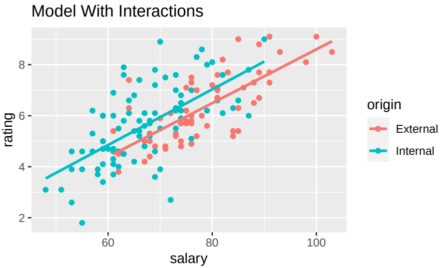
Internal

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	-1.69352	0.94925	-1.784	0.0779
salary	0.10909	0.01407	7.756	1.65e-11

External

	Estimate	Std. Error	t value	$Pr(> \|t\|)$
(Intercept)	-1.9369	0.9862	-1.964	0.0542
salary	0.1054	0.0125	8.432	9.01e-12

Separate Regressions of Manager Rating on Salary



Internal

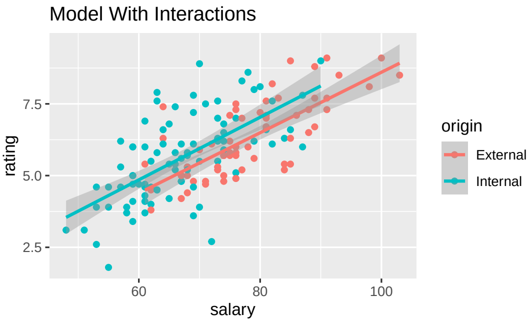
	Estimate	Std. Error	t value	$Pr(>|t|)$
(Intercept)	-1.69352	0.94925	-1.784	0.0779
salary	0.10909	0.01407	7.756	1.65e-11

External

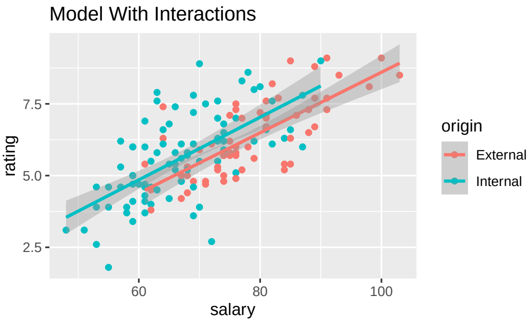
	Estimate	Std. Error	t value	$Pr(>|t|)$
(Intercept)	-1.9369	0.9862	-1.964	0.0542
salary	0.1054	0.0125	8.432	9.01e-12

- At any given salary, internal managers get higher average ratings!
- In regression, *confounding* is a form of *collinearity*.
  - *Salary* is related to *Origin* which was the variable used to explain *Rating*.
  - With *Salary* added, the effect of *Origin* changes sign. Now internal managers look better.

Are the Two Fits Significantly Different?



Are the Two Fits Significantly Different?



- The two confidence bands overlap, which make the comparison indecisive.
- A more powerful idea is to combine these two separate simple regressions into one multiple regression that will allow us to compare these fits.

Regress Manager Rating on both Salary and Origin

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.100459	0.768140	-2.734	0.00702
originInternal	0.514966	0.209029	2.464	0.01491
salary	0.107478	0.009649	11.139	$<$ 2e-16

- x_1 dummy variable of being 'Internal', $I(Origin = Internal)$
- Notice that we only require one dummy variable to distinguish internal from external managers.

Regress Manager Rating on both Salary and Origin

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.100459	0.768140	-2.734	0.00702
originInternal	0.514966	0.209029	2.464	0.01491
salary	0.107478	0.009649	11.139	$<$ 2e-16

- x_1 dummy variable of being 'Internal', $I(Origin = Internal)$
- Notice that we only require one dummy variable to distinguish internal from external managers.
- This enables two *parallel* lines for two kinds of managers.
  - Origin = External
    Manager Rating = -2.100459 + 0.107478 Salary
  - Origin = Internal
    Manager Rating = -2.100459 + 0.107478 Salary + 0.514966
- The coefficient of the dummy variable is the difference between the intercepts.

Model with Parallel Lines

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-2.100459	0.768140	-2.734	0.00702
originInternal	0.514966	0.209029	2.464	0.01491
salary	0.107478	0.009649	11.139	$<$ 2e-16

- The difference between the intercepts is significantly different from 0, since 0.0149, the p-value for Origin[Internal], is less than 0.05.
- Thus, if we assume the slopes are equal, a model using a categorical predictor implies that *controlling* for initial salary, internal managers rate significantly higher.
- How can we check the assumption that the slopes are parallel?

Model with Interaction: Different Slopes

- Beyond just looking at the plot, we can fit a model that allows the slopes to differ.
- This model gives an estimate of the difference between the slopes.
- This estimate is known as an *interaction*.
- An interaction between a dummy variable and a numerical variable measures the difference between the slopes of the numerical variable in the two groups.

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.936941	1.156482	-1.675	0.0961
originInternal	0.243417	1.447230	0.168	0.8667
salary	0.105391	0.014657	7.191	3.09e-11
originInternal:salary	0.003702	0.019520	0.190	0.8499

- *Interaction* variable – product of the dummy variable and Salary:

originInternal:salary	=salary	if Origin = Internal
	=0	if Origin = External

- Origin = External
  Manager Rating = -1.94 + 0.11 Salary
- Origin = Internal
  Manager Rating = (-1.94+0.24) + (0.11+0.0037) Salary
  = -1.69 + 0.11 Salary

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-1.936941	1.156482	-1.675	0.0961
originInternal	0.243417	1.447230	0.168	0.8667
salary	0.105391	0.014657	7.191	3.09e-11
originInternal:salary	0.003702	0.019520	0.190	0.8499

- *Interaction* variable – product of the dummy variable and Salary:

originInternal:salary	=salary	if Origin = Internal
	=0	if Origin = External

- Origin = External
  Manager Rating = -1.94 + 0.11 Salary
- Origin = Internal
  Manager Rating = (-1.94+0.24) + (0.11+0.0037) Salary
  = -1.69 + 0.11 Salary
- These equations *match* the simple regressions fit to the two groups separately.
  The interaction is *not significant* because its *p*-value is large.

Principle of Marginality

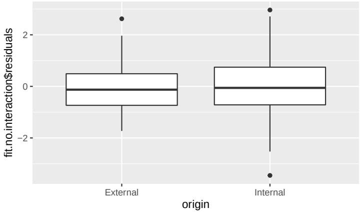
- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.

Principle of Marginality

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.
- *Origin* became insignificant when *Salary*∗*Origin* was added, which is due to collinearity.

Principle of Marginality

- Leave *main effects* in the model (here *Salary* and *Origin*) whenever an interaction that uses them is present in the fitted model. If the interaction is not statistically significant, remove the interaction from the model.
- *Origin* became insignificant when *Salary∗Origin* was added, which is due to collinearity.
- The assumption of equal error variance should also be checked by comparing boxplots of the residuals grouped by the levels of the categorical variable.



Summary

- Categorical variables model the differences between groups using regression, while taking account of other variables.
- In a model with a categorical variable, the *coefficients of the categorical terms* indicate *differences between parallel lines*.
- In a model that includes interactions, the *coefficients of the interaction* measure the *differences in the slopes* between the groups.
- Significant categorical variable $\Rightarrow$ different intercepts
- Significant interaction $\Rightarrow$ different slopes

Further Study

- Reading: Ch 3, An Introduction to Statistical Learning with Applications in R by James et al.
- Model selection, Classification, Nonlinear (trees, random forest, SVM, boosting, deep learning), Unsupervised (clustering, PCA)

Further Study

- What if conditions of regressions are violated?
  - Linear relationship $\rightarrow$ Nonlinear models (transformations, generalized linear models, non-parametric methods, machine learning techniques)
  - Normal residuals $\rightarrow$ Likelihood-based approach
  - Independent residuals $\rightarrow$ Instrumental variable
  - Constant variability $\rightarrow$ Generalized least squares
  - No extreme outliers $\rightarrow$ Robust statistics
  - No strong collinearity $\rightarrow$ Bias-variance tradeoff, Penalized regression (ridge regression, LASSO, SCAD etc)
  - Low-dimension $K \ll n$ $\rightarrow$ High-dimensional inference