Model Overview

So Many Models

- Someone who appears in magazines?
- A small car or railroad?
- A data model (such as an ERD)?
- An “object” created in R that we can use for data understanding and data prediction

Why Use a Linear Model?

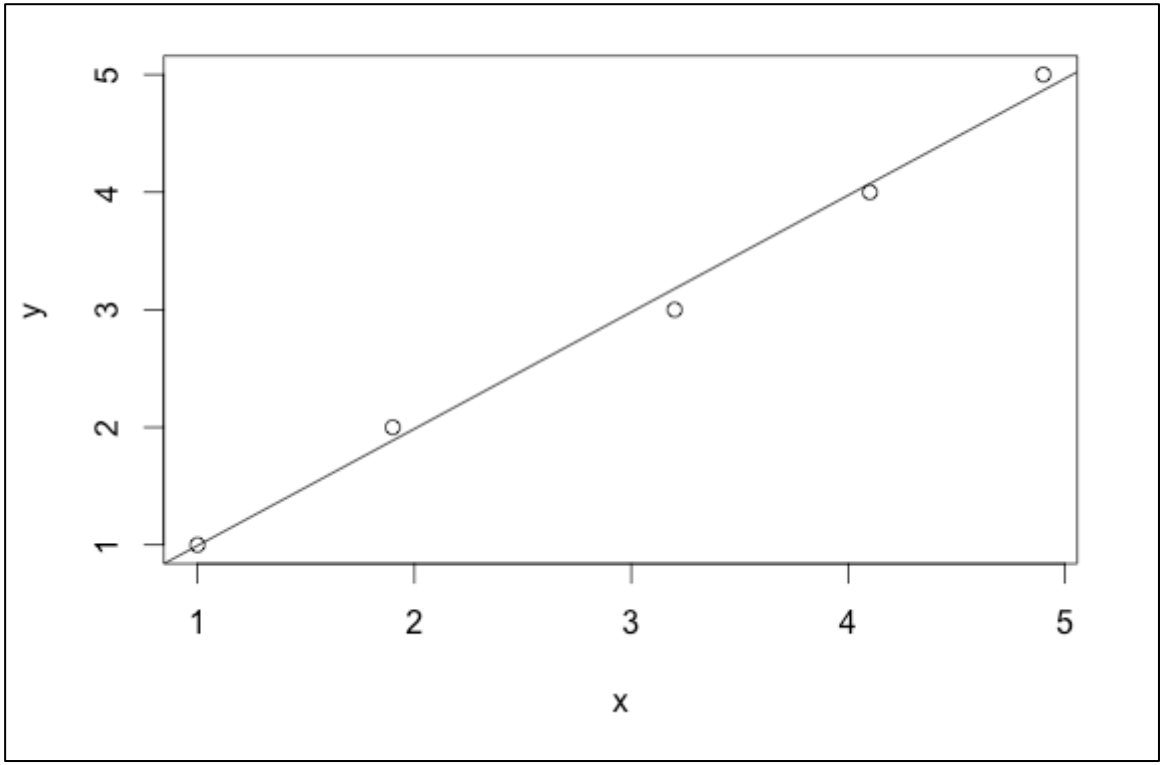
A linear model

—used for prediction (kind of like extrapolation)

Note that:

- It's not "perfect"
- Minimize "distance" from points to the line

$Y = MX + B$



An Example

- Car maintenance (how often to change the oil)
- We manage a "fleet" of cars
  - Cars get replaced every three years
  - Have information on:
    - Past repairs
    - Miles driven
    - # of oil changes during past three years

→ Can we build a model to predict the cost of repairs?

Data for the Analysis

	oilChanges	repairs	miles
1	3	300	20100
2	5	300	23200
3	2	500	19200
4	3	400	22100
5	1	700	18400
6	4	420	23400
7	6	100	17900
8	4	290	19900
9	3	475	20100
10	2	620	24100
11	0	600	18200
12	10	0	19600
13	7	200	20800
14	8	50	19700

Question

Which are independent and which are dependent variables?

Why?

	oilChanges	repairs	miles
1	3	300	20100
2	5	300	23200
3	2	500	19200
4	3	400	22100
5	1	700	18400
6	4	420	23400
7	6	100	17900
8	4	290	19900
9	3	475	20100
10	2	620	24100
11	0	600	18200
12	10	0	19600
13	7	200	20800
14	8	50	19700

Working Through an Example

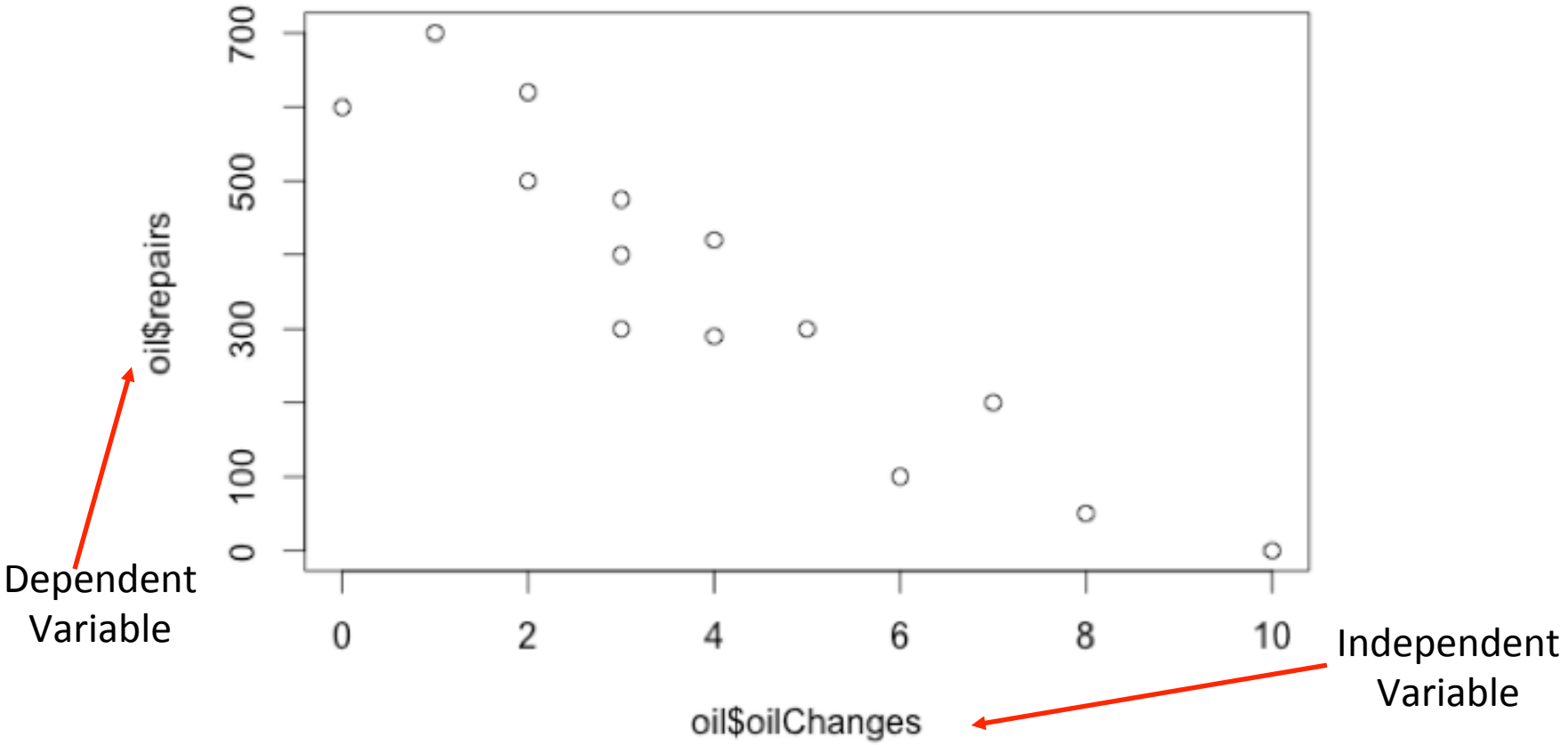
Back to our Data

Let's use this data to build a model

	oilChanges	repairs	miles
1	3	300	20100
2	5	300	23200
3	2	500	19200
4	3	400	22100
5	1	700	18400
6	4	420	23400
7	6	100	17900
8	4	290	19900
9	3	475	20100
10	2	620	24100
11	0	600	18200
12	10	0	19600
13	7	200	20800
14	8	50	19700

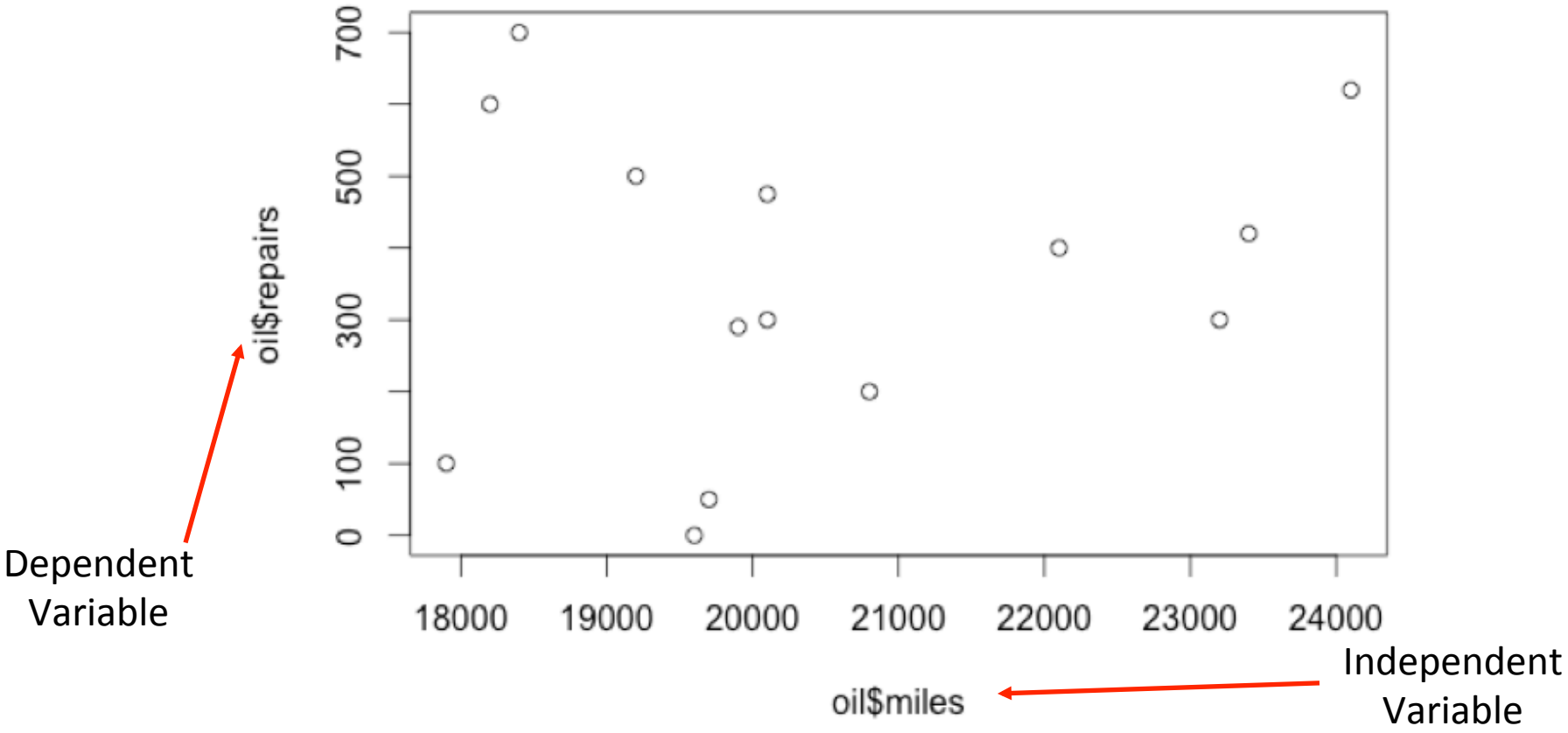
Exploring the Data

> plot(oil$oilChanges, oil$repairs)



Exploring the Data

> plot(oil$miles, oil$repairs)



Generating the First Model

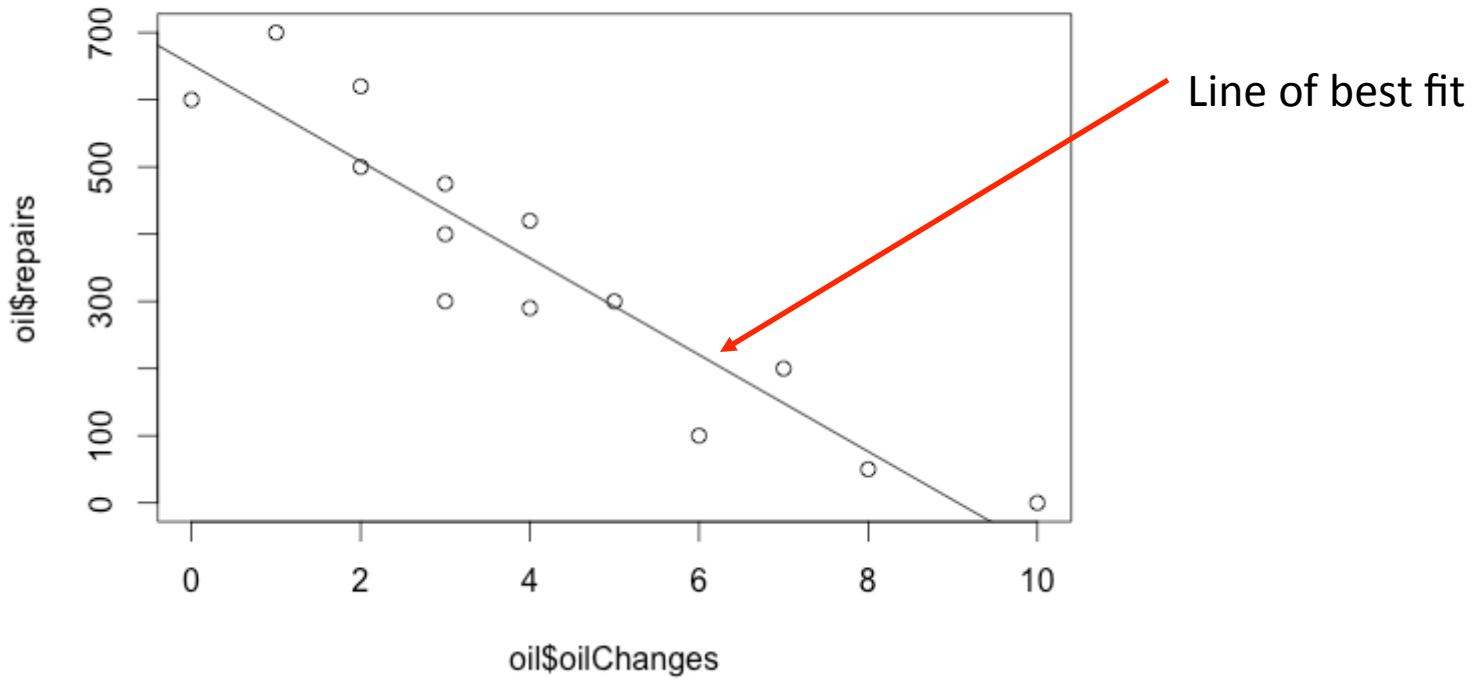
```
> model1 <- lm(formula=repairs ~ oilChanges, data=oil)
> summary(model1)

Call:
lm(formula = repairs ~ oilChanges, data = oil)

Residuals:
    Min      1Q  Median      3Q     Max
-136.208  -48.195   -0.211   54.782  119.803

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  652.191     40.537  16.089 1.74e-09 ***
oilChanges   -71.994      8.202  -8.778 1.44e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 82.72 on 12 degrees of freedom
Multiple R-squared:  0.8653,    Adjusted R-squared:  0.854
F-statistic: 77.05 on 1 and 12 DF,  p-value: 1.436e-06
```

Interpreting the Model

- R-squared value 0.8653.
- Known as the coefficient of determination
- The proportion of the variation that is accounted for in the dependent variable by the whole set of independent variables.
- The closer to 1.0 , the greater the influence the independent variable has on predicting the value of the dependent variable.
- The R-squared value of 0.8653 indicates that the oil changes accounts for 86.53% of the cost of repairs.

Looking at the “abline”

abline(model1)



The model suggests that we should do as many oil changes as possible.
→ it predicts very low (almost 0) repairs if we do 9 or more oil changes, but about $680 if we do no oil changes.

Question

What if we factor in the cost to change the oil?

→ How “model” the cost?

→ What might be some ranges of the cost?

Working Through a Refined Example

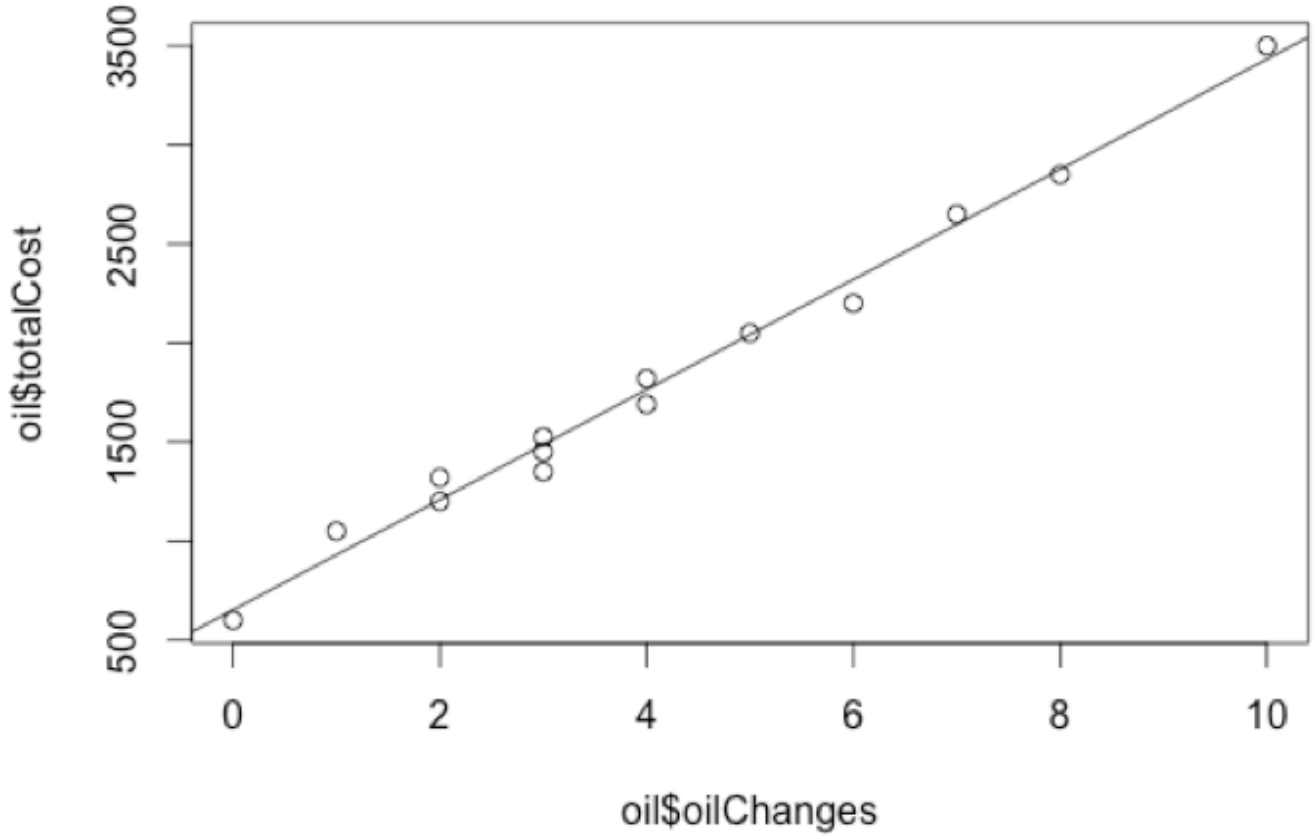
Include the Cost of an Oil Change

- What if oil changes cost $350 each?

```
> oil$oilChangeCost <- oil$oilChanges * 350
> oil$totalCost <- oil$oilChangeCost +
                                        oil$repairs
> m <- lm(formula=totalCost ~ oilChanges,
                                          data=oil)
> plot(oil$oilChanges, oil$totalCost)
> abline(m)
```

Viewing the Data

- What if oil changes cost $350 each?



Using the Model to Predict

- Prediction equation

```
> test = data.frame(oilChanges=0)
> predict(m,test, type="response")
652.191

> test = data.frame(oilChanges=5)
> predict(m,test, type="response")
2042.219

> test = data.frame(oilChanges=10)
> predict(m,test, type="response")
3432.247
```

Question

How accurate is the model?

- Did we have all the facts?
- Did we have all the data?


SYRACUSE UNIVERSITY
SUOS
CULTORES
SCIENTIA
CORONAT