COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

Outline

- The Linear Regression Model
    - Least Squares Fit
    - Measures of Fit
    - Inference in Regression
- Other Considerations in Regression Model
    - Qualitative Predictors
    - Interaction Terms
- Potential Fit Problems

The Linear Regression Model

$$\boxed{Y_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e}$$

- The parameters in the linear regression model are very easy to interpret.

- β_0 is the intercept (i.e. the average value for Y if all the X's are zero), β_j is the slope for the jth variable X_j

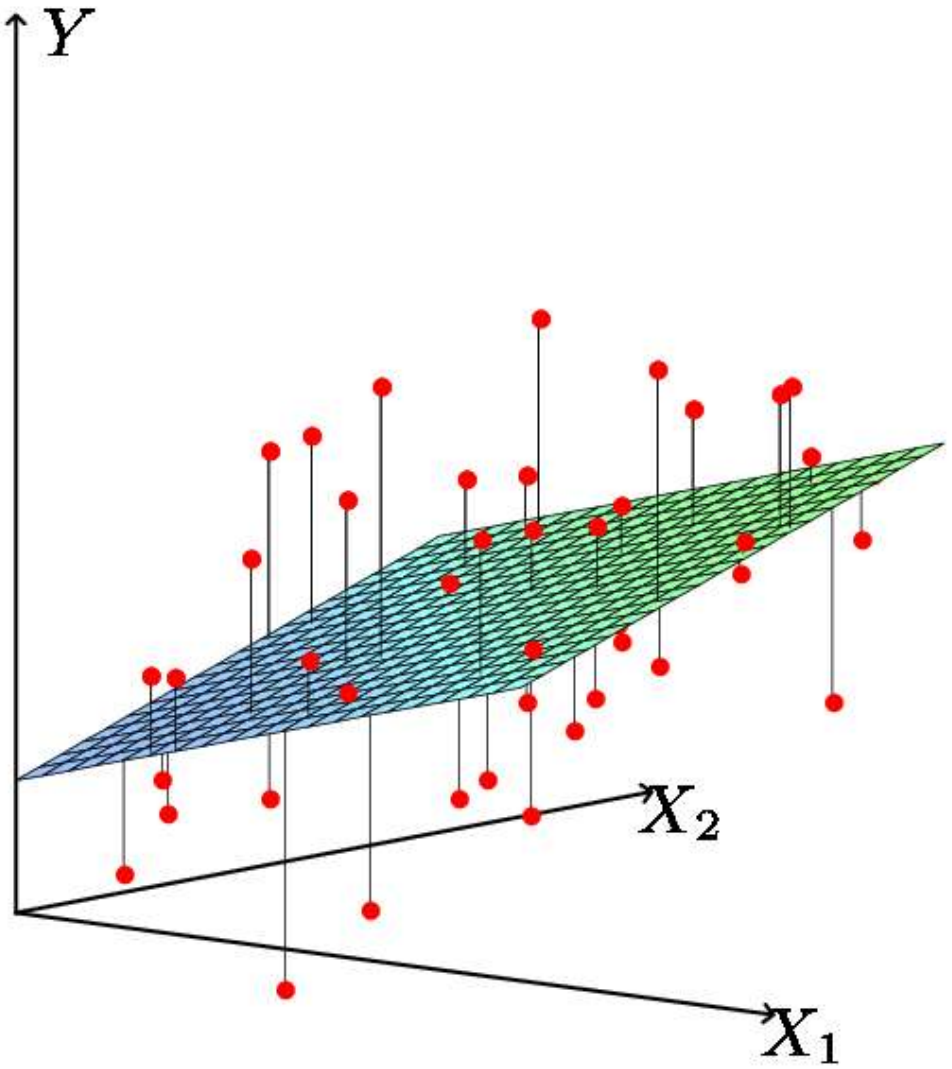
- β_j is the average increase in Y when X_j is increased by one and **all other X's are held constant.**

Least Squares Fit

- We estimate the parameters using least squares i.e. minimize

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \cdots - \hat{\beta}_p X_p\right)^2$$



Relationship between population and least squares lines

Population line

$$\boxed{Y_i = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p + e}$$

Least Squares line

$$\boxed{\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 + \cdots + \hat{b}_p X_p}$$

- We would like to know β_0 through β_p i.e. the population line. Instead we know $\hat{\beta}_0$ through $\hat{\beta}_p$ i.e. the least squares line.
- Hence we use $\hat{\beta}_0$ through $\hat{\beta}_p$ as guesses for β_0 through β_p and $\hat{Y}_i$ as a guess for Y_i. The guesses will not be perfect just as $\bar{X}$ is not a perfect guess for μ.

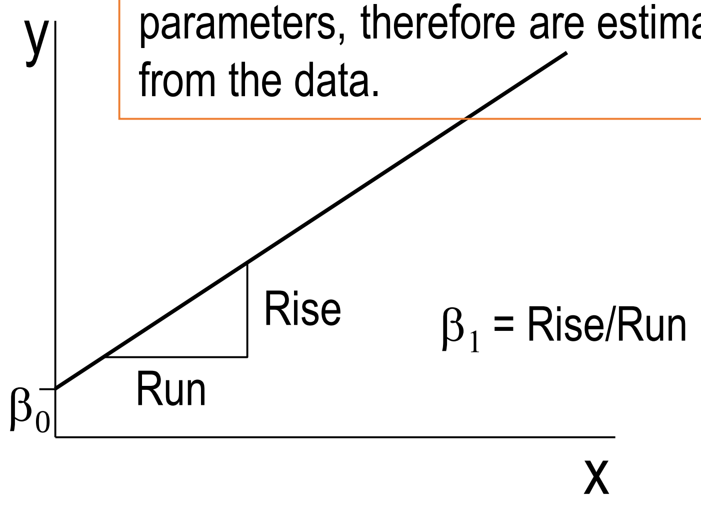
The Model

- The first order linear model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

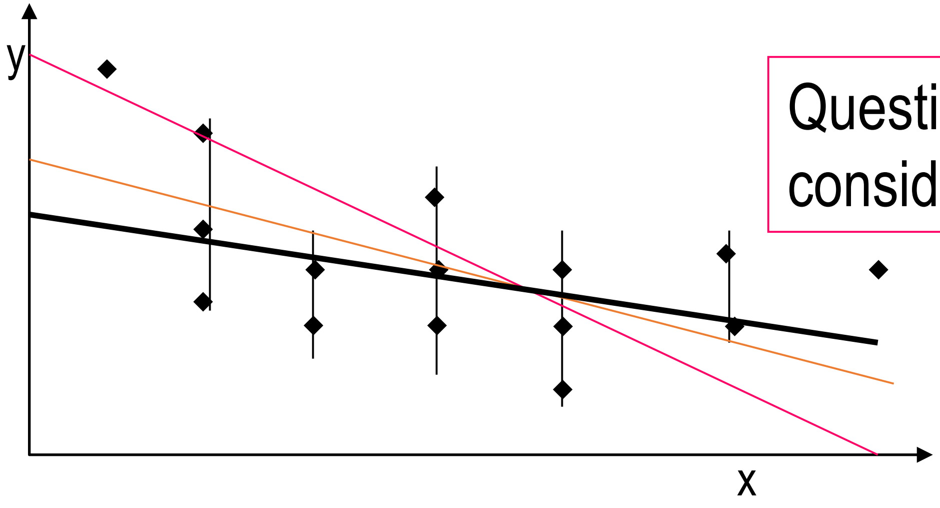
y = dependent variable
x = independent variable
β_0 = y-intercept
β_1 = slope of the line
ε = error variable

β_0 and β_1 are unknown population parameters, therefore are estimated from the data.



Estimating the Coefficients

- The estimates are determined by
  - drawing a sample from the population of interest,
  - calculating sample statistics.
  - producing a straight line that cuts into the data.



Question: What should be considered a good line?

The Least Squares (Regression) Line

A good line is one that minimizes
the sum of squared differences between the points and the line.

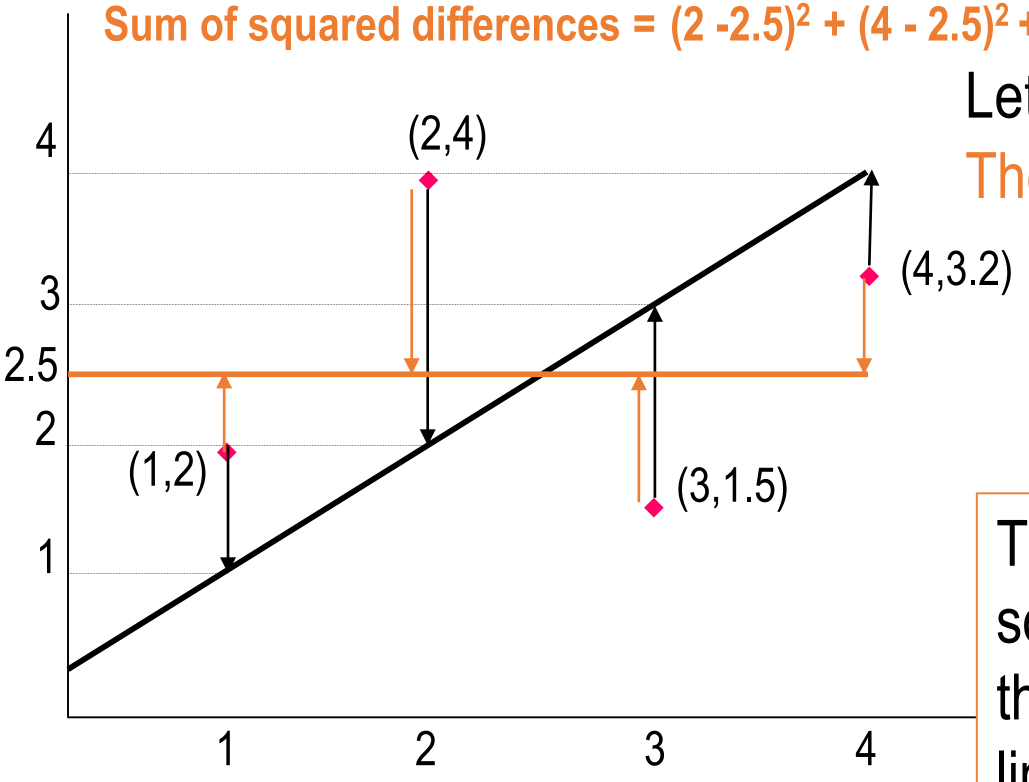
The Least Squares (Regression) Line

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$

Let us compare two lines

The second line is horizontal



The smaller the sum of squared differences the better the fit of the line to the data.

The Estimated Coefficients

To calculate the estimates of the slope and intercept of the least squares line , use the formulas:

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = (n-1)s_x^2$$

Alternate formula for the slope b_1

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

The regression equation that estimates the equation of the first order linear model is:

$$\hat{y} = b_0 + b_1 x$$

The Simple Linear Regression Line

- Example:
  - A car dealer wants to find the relationship between the odometer reading and the selling price of used cars.
  - A random sample of 100 cars is selected, and the data recorded.
  - Find the regression line.

Car	Odometer	Price
1	37388	14636
2	44758	14122
3	45833	14016
4	30862	15590
5	31705	15568
6	34010	14718
.	.	.
.	.	.
.	.	.

Independent variable x | Dependent variable y

The Simple Linear Regression Line

- Solution
  - Solving by hand: Calculate a number of statistics

$$\bar{x} = 36{,}009.45; \qquad SS_{xx} = \sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n} = 43{,}528{,}690$$

$$\bar{y} = 14{,}822.823; \qquad SS_{xy} = \sum (x_i y_i) - \frac{\sum x_i \sum y_i}{n} = -2{,}712{,}511$$

where n = 100.

$$b_1 = \frac{SS_{xy}}{(n-1)s_x^2} = \frac{-2{,}712{,}511}{43{,}528{,}690} = -.06232$$

$$b_0 = \bar{y} - b_1\bar{x} = 14{,}822.82 - (-.06232)(36{,}009.45) = 17{,}067$$

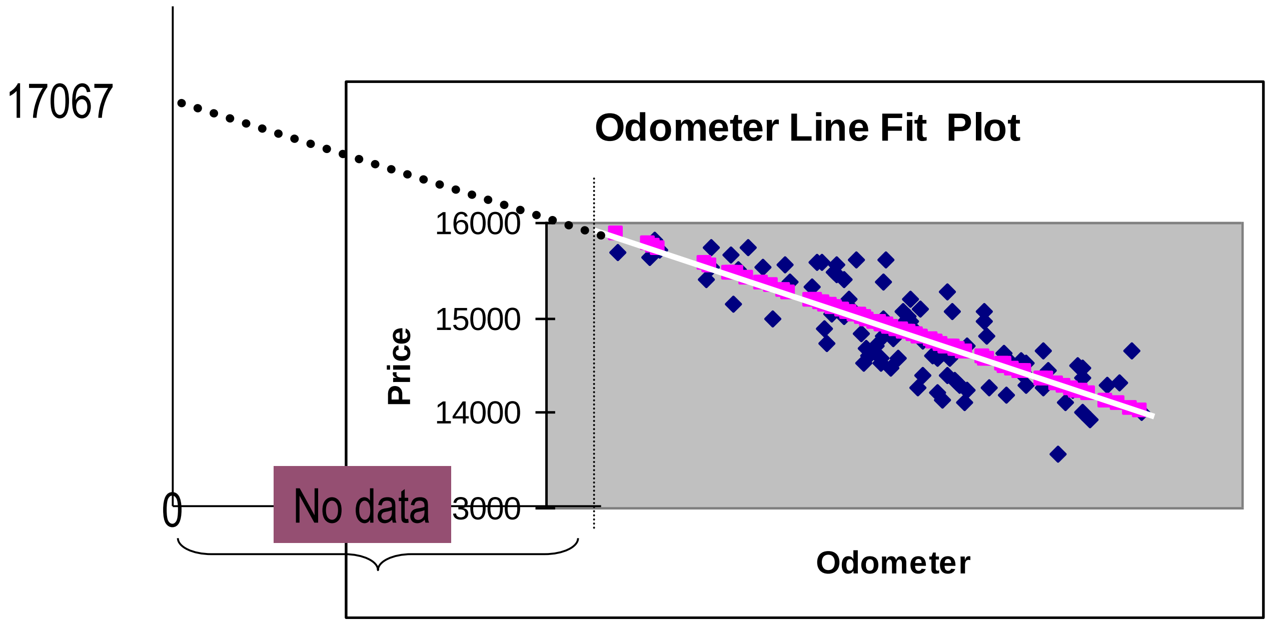
$$\boxed{\hat{y} = b_0 + b_1 x = 17{,}067 - .0623x}$$

The Simple Linear Regression Line

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.8063					
R Square	0.6501					
Adjusted R	0.6466					
Standard E	303.1					
Observatio	100					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	16734111	16734111	182.11	0.0000	
Residual	98	9005450	91892			
Total	99	25739561				
	Coefficients	*Standard Erro*	*t Stat*	*P-value*		
Intercept	17067	169	100.97	0.0000		
Odometer	-0.0623	0.0046	-13.49	0.0000		

$$\hat{y} = 17{,}067 - .0623\,x$$

Interpreting the Linear Regression -Equation



$$\hat{y} = 17,067 - .0623\,x$$

The intercept is b_0 = \$17067.

This is the slope of the line.
For each additional mile on the odometer,
the price decreases by an average of \$0.0623

Do not interpret the intercept as the
"Price of cars that have not been driven"

Coefficient of determination

- To measure the strength of the linear relationship we use the coefficient of determination.

$$R^2 = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{s_x^2 s_y^2}$$

$$or \quad R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$$

Note that the coefficient of determination is r^2

Coefficient of determination

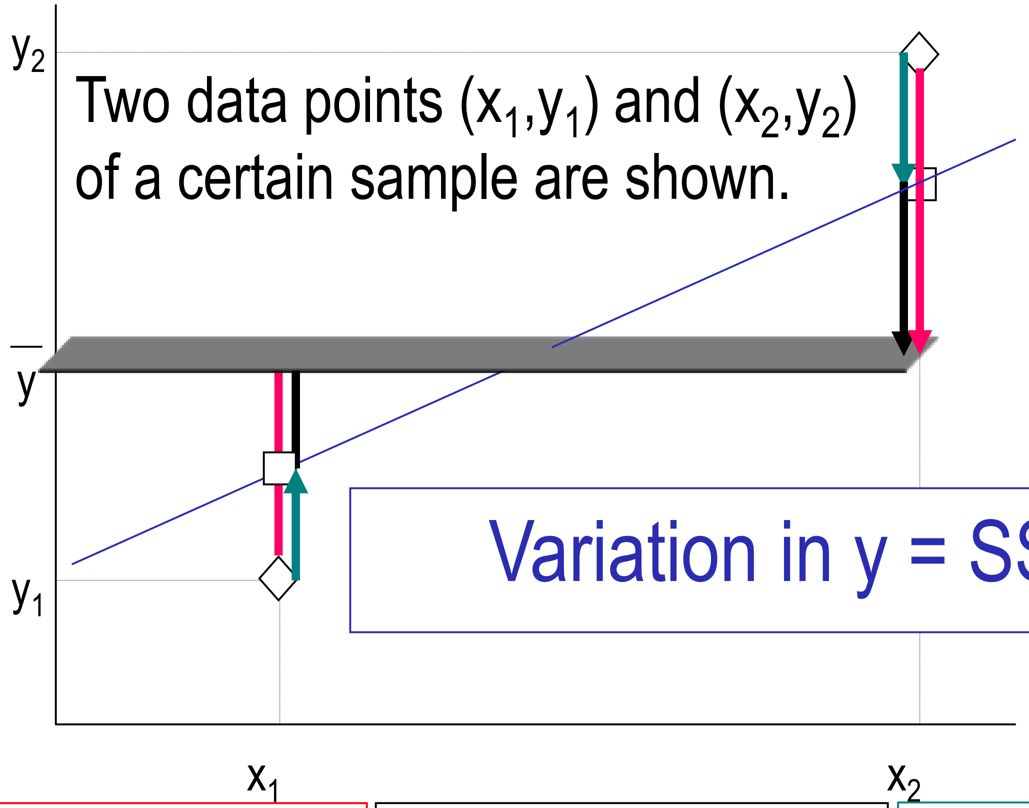
- To understand the significance of this coefficient note:

Overall variability in y

Explained in part by → The regression model

Remains, in part, unexplained → The error

Coefficient of determination



Variation in y = SSR + SSE

Total variation in y = | Variation explained by the regression line | + Unexplained variation (error)

$$(y_1-\bar{y})^2+(y_2-\bar{y})^2 = (\hat{y}_1-\bar{y})^2+(\hat{y}_2-\bar{y})^2 + (y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2$$

Coefficient of determination

- R^2 measures the proportion of the variation in y that is explained by the variation in x.

$$R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - SSE}{\sum (y_i - \bar{y})^2} = \frac{SSR}{\sum (y_i - \bar{y})^2}$$

- R^2 takes on any value between zero and one.

  $R^2 = 1$: Perfect match between the line and the data points.

  $R^2 = 0$: There are no linear relationship between x and y.

Coefficient of determination

- Example
  - Find the coefficient of determination for the used car price –odometer example. What does this statistic tell you about the model?
- Solution
  - Solving by hand;

$$R^2 = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{s_x^2 s_y^2} = \frac{[-2{,}712{,}511]^2}{(43{,}528{,}688)(259{,}996)} = .6501$$

Coefficient of determination

– Using Excel

From the regression output we have

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8063
R Square	0.6501
Adjusted R S	0.6466
Standard Err	303.1
Observations	100

ANOVA

	df	SS	MS	F	
Regression	1	16734111	16734111	182.11	0.0000
Residual	98	9005450	91892		
Total	99	25739561			

	Coefficients	Standard Erro	t Stat	P-value
Intercept	17067	169	100.97	0.0000
Odometer	-0.0623	0.0046	-13.49	0.0000

65% of the variation in the auction selling price is explained by the variation in odometer reading. The rest (35%) remains unexplained by this model.

Measures of Fit: R^2

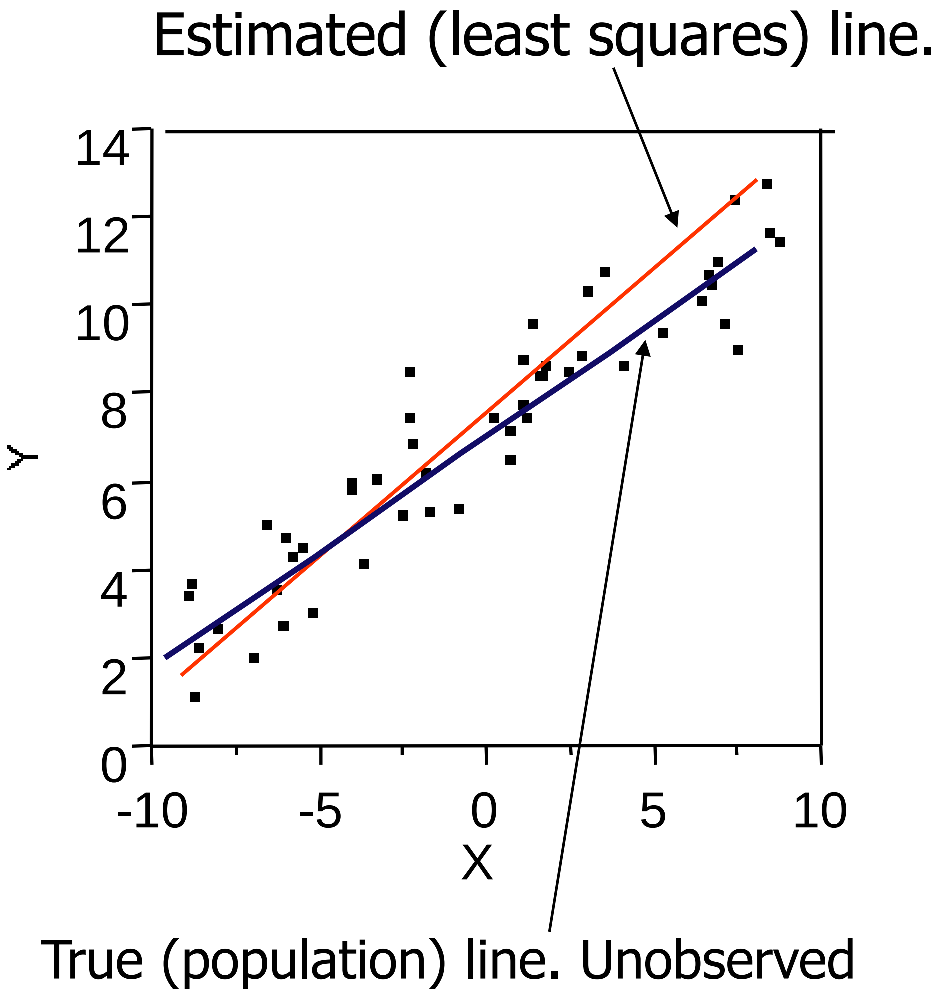
- Some of the variation in Y can be explained by variation in the X's and some cannot.
- R^2 tells you the fraction of variance that can be explained by X.

$$R^2 = 1 - \frac{RSS}{\sum (Y_i - \bar{Y})^2} \approx 1 - \frac{\text{Ending Variance}}{\text{Starting Variance}}$$

R^2 is always between 0 and 1. Zero means no variance has been explained. One means it has all been explained (perfect fit to the data).

Inference in Regression

- The regression line from the sample is not the regression line from the population.
- What we want to do:
  - Assess how well the line describes the plot.
  - Guess the slope of the population line.
  - Guess what value Y would take for a given X value



Assessing the Model

- The least squares method will produces a regression line whether or not there is a linear relationship between x and y.
- Consequently, it is important to assess how well the linear model fits the data.
- Several methods are used to assess the model. All are based on the sum of squares for errors, SSE.

Sum of Squares for Errors

- This is the sum of differences between the points and the regression line.
- It can serve as a measure of how well the line fits the data. SSE is defined by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

  - A shortcut formula

$$SSE = \sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i$$

Standard Error of Estimate

- The mean error is equal to zero.
- If σ_ε is small the errors tend to be close to zero (close to the mean error). Then, the model fits the data well.
- Therefore, we can, use σ_ε as a measure of the suitability of using a linear model.
- An estimator of σ_ε is given by s_ε

$$Standard\ Error\ of\ Estimate$$

$$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$$

Standard Error of Estimate

- Example:
  - Calculate the standard error of estimate for the previous example and describe what it tells you about the model fit.
- Solution

$$SSE = 9,005,450$$

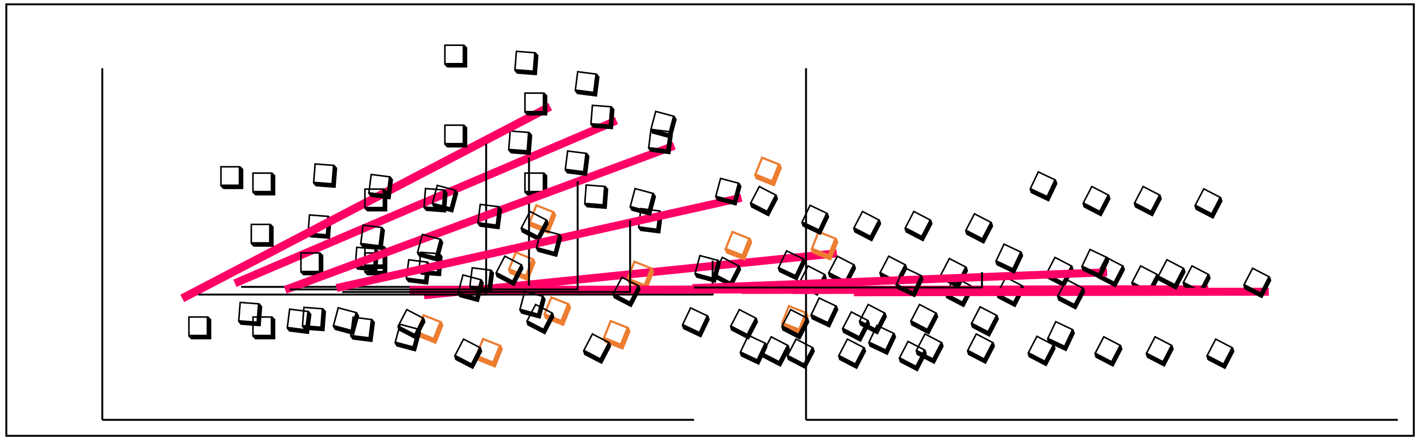
$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{9,005,450}{98}} = 303.13$$

It is hard to assess the model based on s_{ε} even when compared with the mean value of y.

$s_{\varepsilon} = 303.1 \quad \bar{y} = 14,823$

Testing the slope

- When no linear relationship exists between two variables, the regression line should be horizontal.



Linear relationship.
Different inputs (x) yield different outputs (y).

The slope is not equal to zero

No linear relationship.
Different inputs (x) yield the same output (y).

The slope is equal to zero

Testing the Slope

- We can draw inference about β_1 from b_1 by testing

  H_0: $\beta_1 = 0$

  H_1: $\beta_1 = 0$ (or < 0,or > 0)

  - The test statistic is

$$t = \frac{b_1 - \beta_1}{s_{b_1}} \quad \text{where} \quad s_{b_1} = \frac{s_\varepsilon}{\sqrt{SS_{xx}}}$$

The standard error of b_1.

  - If the error variable is normally distributed, the statistic is Student t distribution with d.f. = n-2.

Testing the Slope

- Example
  - Test to determine whether there is enough evidence to infer that there is a linear relationship between the car auction price and the odometer reading for all three-year-old Tauruses in the previous example .
    Use α = 5%.

Testing the Slope

- Solving by hand
  - To compute "t" we need the values of b_1 and s_{b1}.

$$b_1 = -.0623$$

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}} = \frac{303.1}{\sqrt{(99)(43,528,690)}} = .00462$$

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{-.0623 - 0}{.00462} = -13.49$$

  - The rejection region is $t > t_{.025}$ or $t < -t_{.025}$ with $\nu = n-2 = 98$.
    Approximately, $t_{.025} = 1.984$

Testing the Slope

- Using Excel

Price	Odometer	SUMMARY OUTPUT				
14636	37388					
14122	44758	*Regression Statistics*				
14016	45833	Multiple R	0.8063			
15590	30862	R Square	0.6501			
15568	31705	Adjusted R S	0.6466			
14718	34010	Standard Err	303.1			
14470	45854	Observations	100			
15690	19057					
15072	40149	ANOVA				
14802	40237		*df*	*SS*		
15190	32359	Regression	1	16734111	1	
14660	43533	Residual	98	9005450		
15612	32744	Total	99	25739561		
15610	34470					
14634	37720		*Coefficients*	*Standard Err*	*t Stat*	*P-value*
14632	41350	Intercept	17067	169	100.97	0.0000
15740	24469	Odometer	-0.0623	0.0046	-13.49	0.0000

There is overwhelming evidence to infer that the odometer reading affects the auction selling price.

Some Relevant Questions

1. Is β_j=0 or not? We can use a hypothesis test to answer this question. If we can't be sure that $\beta_j \neq 0$ then there is no point in using X_j as one of our predictors.

1. Can we be sure that at least one of our X variables is a useful predictor i.e. is it the case that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$?

1. Is β_j=0 i.e. is X_j an important variable?

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

$$t = \hat{\beta}_j \big/ SE(\hat{\beta}_j)$$

Number of standard deviations away from zero.

- If t is large (equivalently p-value is small) we can be sure that $\beta_j \neq 0$ and that there is a relationship

Regression coefficients				
	Coefficient	Std Err	t-value	p-value
Constant	7.0326	0.4578	15.3603	0.0000
TV	0.0475	0.0027	17.6676	0.0000

$\hat{\beta}_1$ $SE(\hat{\beta}_1)$

P-value

$\hat{\beta}_1$ is 17.67 SE's from 0

Testing Individual Variables

Is there a (statistically detectable) linear relationship between Newspapers and Sales after all the other variables have been accounted for?

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	2.9389	0.3119	9.4223	0.0000
TV	0.0458	0.0014	32.8086	0.0000
Radio	0.1885	0.0086	21.8935	0.0000
Newspaper	-0.0010	0.0059	-0.1767	0.8599

No: big p-value

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	12.3514	0.6214	19.8761	0.0000
Newspaper	0.0547	0.0166	3.2996	0.0011

Small p-value in simple regression

Almost all the explaining that Newspapers could do in simple regression has already been done by TV and Radio in multiple regression!

2. Is the whole regression explaining anything at all?

- Test for:
  - H_0: all slopes = 0 $(\beta_1=\beta_2=\cdots=\beta_p=0)$,
  - H_a: at least one slope $\neq 0$

ANOVA Table

Source	df	SS	MS	F	p-value
Explained	2	4860.2347	2430.1174	859.6177	0.0000
Unexplained	197	556.9140	2.8270		

Answer comes from the F test in the ANOVA (ANalysis Of VAriance) table.
The ANOVA table has many pieces of information. What we care about is the F Ratio and the corresponding p-value.

Outline

- The Linear Regression Model
  - Least Squares Fit
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems

Qualitative Predictors

- How do you stick "men" and "women" (category listings) into a regression equation?

- Code them as indicator variables (dummy variables)

- For example we can "code" Males=0 and Females= 1.

Interpretation

- Suppose we want to include income and gender.
- Two genders (male and female). Let $\text{Gender}_i = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$
- then the regression equation is $Y_i \approx b_0 + b_1\text{Income}_i + b_2 Gender_i = \begin{cases} b_0 + b_1\text{Income}_i & \text{if male} \\ b_0 + b_1\text{Income}_i + b_2 & \text{if female} \end{cases}$
- β_2 is the average extra balance each month that females have for given income level. Males are the "baseline".

Regression coefficients

	Coefficient	Std Err	t-value	p-value
Constant	233.7663	39.5322	5.9133	0.0000
Income	0.0061	0.0006	10.4372	0.0000
Gender_Female	24.3108	40.8470	0.5952	0.5521

Other Coding Schemes

- There are different ways to code categorical variables.
- Two genders (male and female). Let

$$Gender_i = \begin{cases} -1 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

- then the regression equation is

$$Y_i \approx \beta_0 + \beta_1 \text{Income}_i + \beta_2 Gender_i = \begin{cases} \beta_0 + \beta_1 \text{Income}_i - \beta_2, & \text{if male} \\ \beta_0 + \beta_1 \text{Income}_i + \beta_2, & \text{if female} \end{cases}$$

- β_2 is the average amount that females are above the average, for any given income level. β_2 is also the average amount that males are below the average, for any given income level.

Other Issues Discussed

- Interaction terms
- Non-linear effects
- Multicollinearity
- Model Selection

Interaction

- When the effect on Y of increasing X_1 depends on another X_2.

- Example:
  - Maybe the effect on Salary (Y) when increasing Position (X_1) depends on gender (X_2)?
  - For example maybe Male salaries go up faster (or slower) than Females as they get promoted.

- Advertising example:
  - TV and radio advertising both increase sales.
  - Perhaps spending money on both of them may increase sales more than spending the same amount on one alone?

Interaction in advertising

$$\boxed{Sales = b_0 + b_1 \times TV + b_2 \times Radio + b_3 \times TV \times Radio}$$

Interaction Term

$$\boxed{Sales = b_0 + (b_1 + b_3 \times Radio) \times TV + b_2 \times Radio}$$

- Spending \$1 extra on TV increases average sales by 0.0191 + 0.0011Radio

$$\boxed{Sales = b_0 + (b_2 + b_3 \times TV) \times Radio + b_2 \times TV}$$

- Spending \$1 extra on Radio increases average sales by 0.0289 + 0.0011TV

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	6.7502202	0.247871	27.23	<.0001*
TV	0.0191011	0.001504	12.70	<.0001*
Radio	0.0288603	0.008905	3.24	0.0014*
TV*Radio	0.0010865	5.242e-5	20.73	<.0001*

Parallel Regression Lines

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	112.77039	1.454773	77.52	<.0001
Gender[female]	1.8600957	0.527424	3.53	0.0005
Gender[male]	-1.860096	0.527424	-3.53	0.0005
Position	6.0553559	0.280318	21.60	<.0001

Regression equation

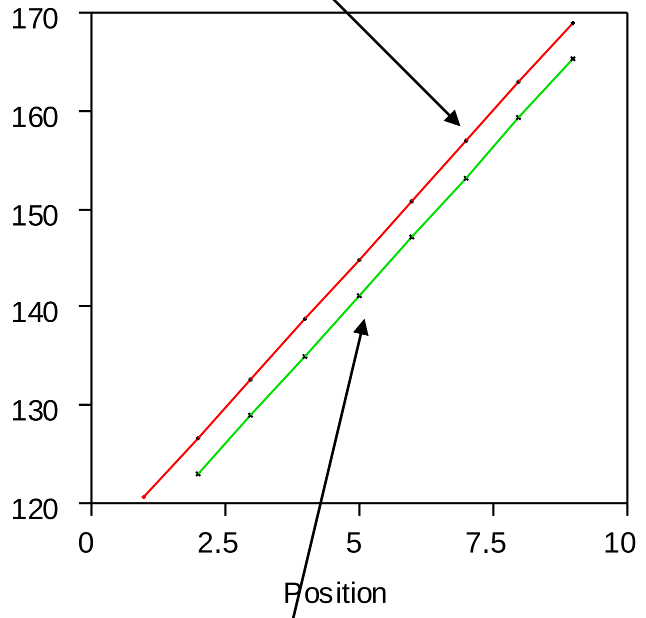
female: salary = 112.77+1.86 + 6.05× position

males: salary = 112.77-1.86 + 6.05 × position

Different intercepts

Same slopes

Line for women



Line for men

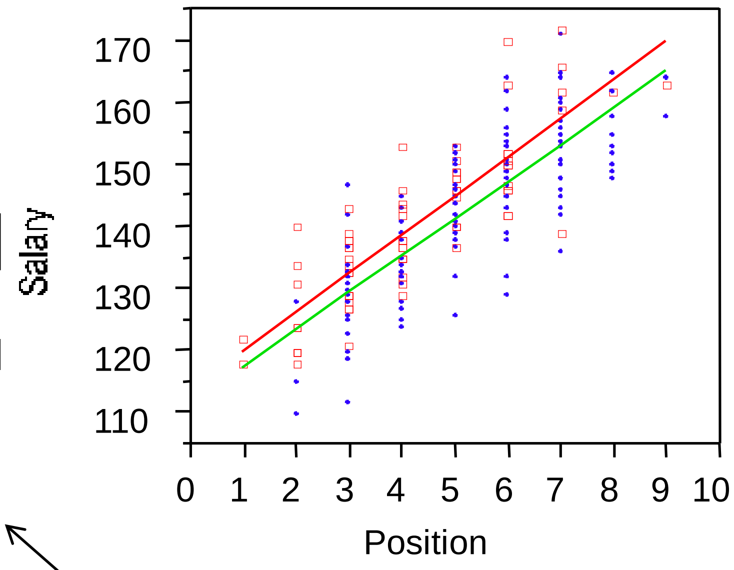
Parallel lines have the same slope. Dummy variables give lines different intercepts, but their slopes are still the same.

Should the Lines be Parallel?

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	112.63081	1.484825	75.85	<.0001
Gender[female]	1.1792165	1.484825	0.79	0.4280
Gender[male]	-1.179216	1.484825	-0.79	0.4280
Position	6.1021378	0.296554	20.58	<.0001
Gender[female]*Position	0.1455111	0.296554	0.49	0.6242
Gender[male]*Position	-0.145511	0.296554	-0.49	0.6242



Interaction is not significant

Interaction between gender and position

Interaction Effects

- Our model has forced the line for men and the line for women to be parallel.

- Parallel lines say that promotions have the same salary benefit for men as for women.

- If lines aren't parallel then promotions affect men's and women's salaries differently.

Extensions of the Linear Model

Removing the additive assumption: *interactions* and *nonlinearity*

Interactions:

- In the analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model

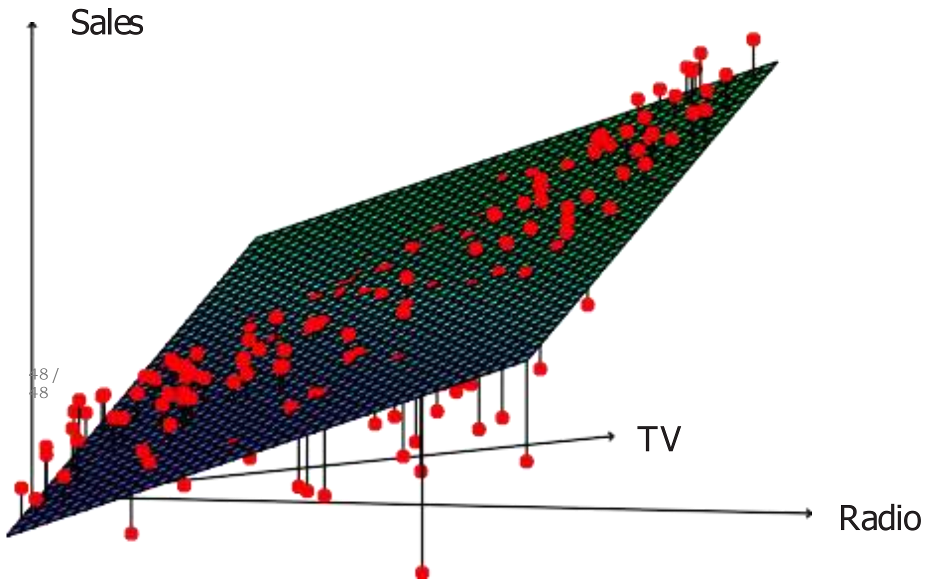
$$\widehat{\text{sales}} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper}$$

states that the average effect on sales of a one-unit increase in TV is always β_1, regardless of the amount spent on radio.

Interactions — continued

- But suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for `TV` should increase as `radio` increases.
- In this situation, given a fixed budget of $100, 000, spending half on `radio` and half on `TV` may increase `sales` more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction* effect.

Interaction in the Advertising data?



When levels of either `TV` or `radio` are low, then the true `sales` are lower than predicted by the linear model.
But when advertising is split between the two media, then the model tends to underestimate `sales`.

Modelling interactions — Advertising data

Model takes the form

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Interpretation

- The results in this table suggests that interactions are important.
- The p-value for the interaction term TV×radio is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.
- The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.

Interpretation — continued

- This means that $(96.8 - 89.7)/(100 - 89.7) =$ 69% of the variability in `sales` that remains after fitting the additive model has been explained by the interaction term.
- The coefficient estimates in the table suggest that an increase in TV advertising of \$1,000 is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units.
- An increase in radio advertising of \$1,000 will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects (in this case, `TV` and `radio`) do not.
- The *hierarchy principle*:

If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

Hierarchy — continued

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model has no main effect terms.

Interactions between qualitative and quantitative variables

Consider the Credit data set, and suppose that we wish to predict balance using income (quantitative) and student (qualitative).

Without an interaction term, the model takes the form

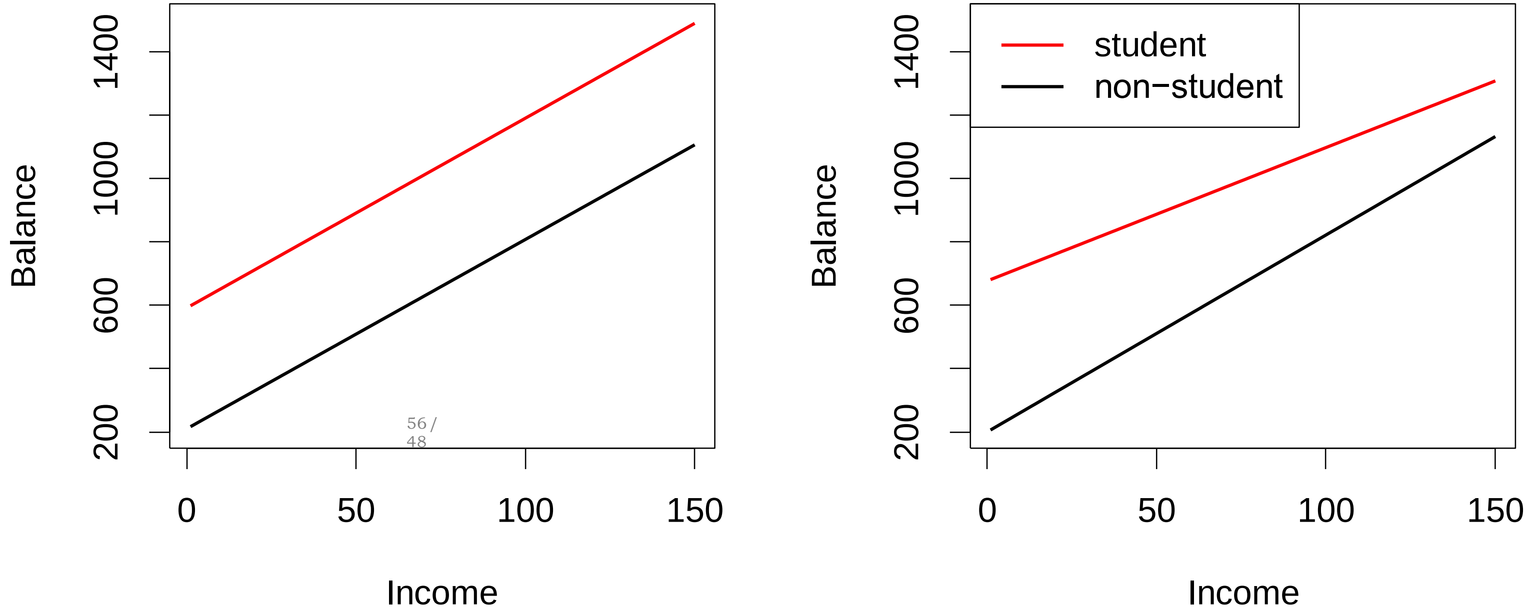
$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}$$

With interactions, it takes the form

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

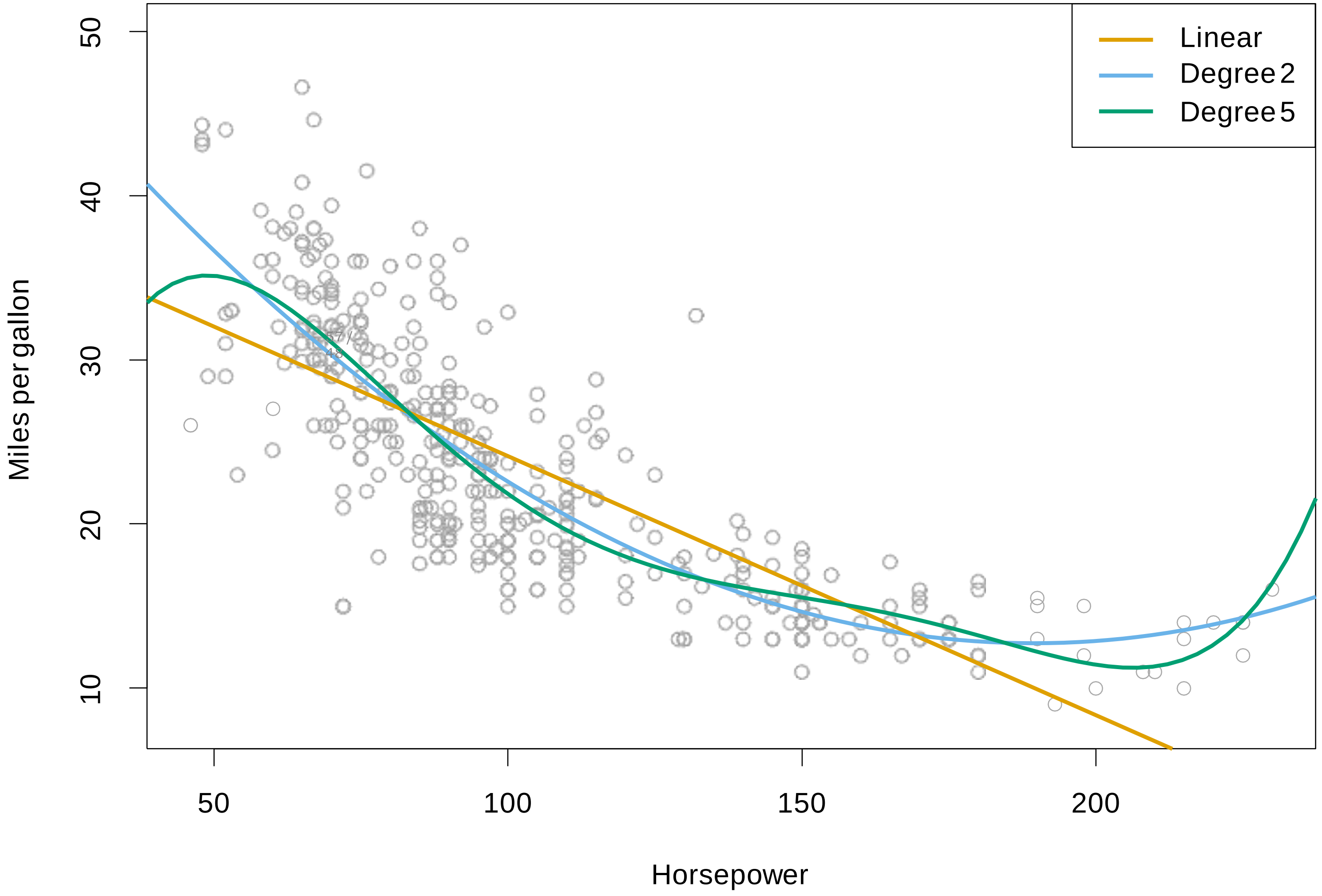
$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if not student} \end{cases}$$



Credit data; Left: no interaction between income and student.
Right: with an interaction term between income and student.

Non-linear effects of predictors

polynomial regression on Auto data



The figure suggests that

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + E$$

may provide a better fit.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

Generalizations of the Linear Model

In much of the rest of this course, we discuss methods that expand the scope of linear models and how they are fit:

- *Classification problems:* logistic regression, support vector machines
- *Non-linearity:* kernel smoothing, splines and generalized additive models; nearest neighbor methods.
- *Interactions:* Tree-based methods, bagging, random forests and boosting (these also capture non-linearities)
- *Regularized fitting:* Ridge regression and lasso

This is the analysis of variance table for a simple linear regression. The "simple" here means that exactly one predictor (call it X) is used to try to explain the dependent variable (call it Y).

The degrees of freedom number in the Regression line is the number of predictors used. Here that number is 1, and indeed the regression has a single predictor.

The F statistic asks whether the regression is statistically significant. As a quick approximation, $F > 4$ suggests that the slope β_1 is not zero.

If the P value is provided, it should be used as the formal test of significance. If $P \le 0.05$, we can say that the sample slope b_1 is significantly different from zero.

```
Analysis of Variance

Source           DF      SS       MS       F      P
Regression        1   835.43   835.43   89.64  0.000
Residual Error   38   354.15     9.32
Total            39  1189.57
```

The total degrees of freedom is $n - 1$, where n is the sample size. Here $n = 40$ data points were used.

Each mean square (MS) is computed as (SS)/(DF). Observe that $9.32 \approx \frac{354.15}{38}$. The MS in the Residual Error line has the added use $\sqrt{MS_{\text{Residual Error}}} = s_\varepsilon$. Here that value is $\sqrt{9.32} \approx 3.0529$. Nearly all software reports this value separately.

By convention, no number is reported in this position. If we followed the (SS)/(DF) pattern, this value would be $\frac{1{,}189.57}{39} \approx 30.5018$. The square root is the sample standard deviation of the dependent variable Y; $s_Y = \sqrt{30.5018} \approx 5.5228$.

You can also find $R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = \frac{835.43}{1{,}189.57} \approx 0.7023$. Nearly all software reports this value separately.

Many programs report also the adjusted R^2, given as $R^2_{\text{adj}} = 1 - \left(\frac{s_\varepsilon}{s_Y}\right)^2 = 1 - \left(\frac{3.0529}{5.5228}\right)^2 \approx 0.6944$. This can also be computed through $R^2_{\text{adj}} = 1 - \frac{n-1}{n-1-K}\left(1-R^2\right)$. Remember that $K = 1$ is the simple regression case.

Outline

- The Linear Regression Model
  - Least Squares Fit
  - Measures of Fit
  - Inference in Regression
- Other Considerations in Regression Model
  - Qualitative Predictors
  - Interaction Terms
- Potential Fit Problems

Potential Fit Problems

There are a number of possible problems that one may encounter when fitting the linear regression model.

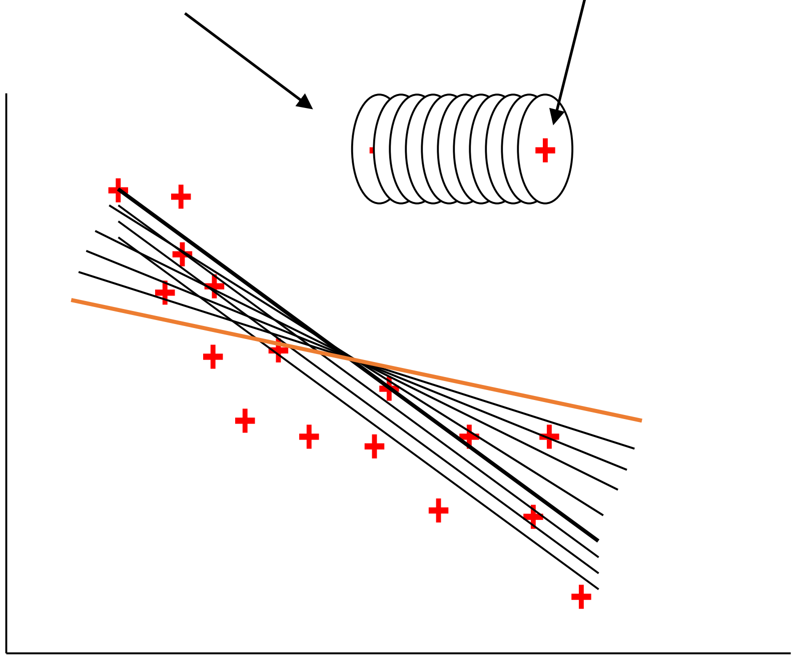
1. Non-linearity of the data
2. Dependence of the error terms
3. Non-constant variance of error terms
4. Outliers
5. High leverage points
6. Collinearity

Outliers

- An outlier is an observation that is unusually small or large.
- Several possibilities need to be investigated when an outlier is observed:
  - There was an error in recording the value.
  - The point does not belong in the sample.
  - The observation is valid.
- Identify outliers from the scatter diagram.
- It is customary to suspect an observation is an outlier if its |standard residual| > 2

An outlier

An influential observation



… but, some outliers may be very influential

The outlier causes a shift in the regression line

Procedure for Regression Diagnostics

- Develop a model that has a theoretical basis.
- Gather data for the two variables in the model.
- Draw the scatter diagram to determine whether a linear model appears to be appropriate.
- Determine the regression equation.
- Check the required conditions for the errors.
- Check the existence of outliers and influential observations
- Assess the model fit.
- If the model fits the data, use the regression equation.