SCS 3251 – Statistics for Data Science

Introduction to Linear Regression

Module 6

INTRODUCTION TO LINEAR REGRESSION

Course Roadmap

Module / Week	Title
1	Introduction to Statistics for Data Science
2	Probability
3	Distribution of Random Variables
4	Inference
5	Model Building
6	Linear Regression
7	Multiple Linear Regression
8	Logistic Regression
9	Introduction to Bayesian Inference
10	Multi-level Models
11	Markov Chain Monte Carlo
12	Presentations
13	Final Exam

Module 6: Learning Objectives

- Apply Ordinary Least Squares Linear Regression
- Use R, Statsmodels and Scikit-Learn

Key Topic Overview – Introduction to Linear Regression

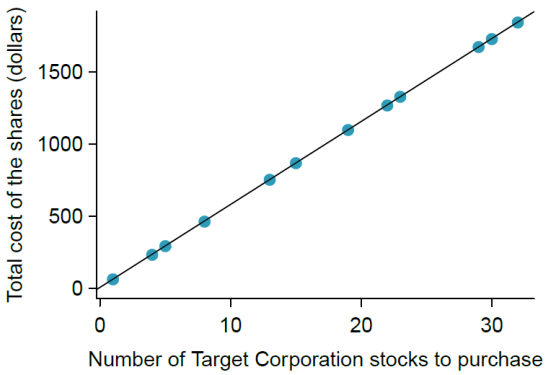
- Concepts, conventions, definitions
- Conditions under which OLS is appropriate
- Interpreting the goodness of fit
- Working with categorical predictors
- Using regression packages

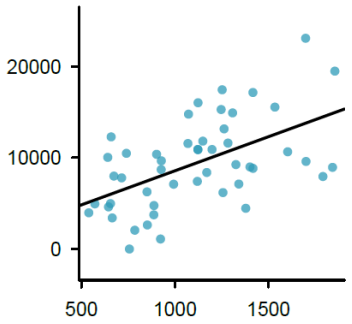
Linear Regression Analysis

- The objective of a regression analysis is to estimate the value of a random variable given that we know the value of an associated variable
- When the association is linear we refer to linear regressions
- A linear regression follows the equation

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

- The strength of a linear relation is quantified by the correlation (defined later)
- However it is unrealistic to expect that you find a perfect linear relationship in any observable process
- In real cases the data will fall ‘around’ a straight line
- Different criteria have been developed to ‘fit’ the line. Because of the variability there is uncertainty associated with the parameters estimated.





Conventions and Definitions

- In the equation $y = \beta_0 + \beta_1 x$ we call
  - x the predictor, explanatory variable or independent variable
  - y the response or dependent variable
  - β_0 the intercept, it indicates the value y when $x = 0$
  - β_1 the slope or coefficient

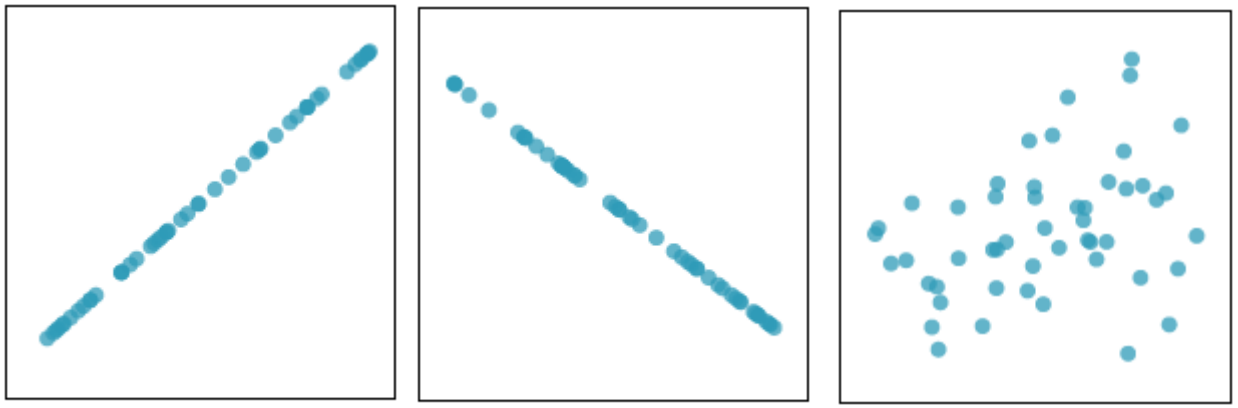
You will see that sometimes the linear relation is expressed as $\hat{y} = \beta_0 + \beta_1 x$. The hat signifies that $\hat{y}$ is an estimate and not the 'real' value. Sometimes… $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

- It is important to note that the fitted value is an estimation that will generally differ from the value observed. The leftover variation after accounting for the model fit is called residual error
- $\text{Data} = \text{Fit} + \text{Residual}$ or $y_i = \hat{y}_i + e_i$

  Where $_i$ denotes the i^{th} observation (y_i, x_i)

Correlation: Strength of a Linear Relationship

Correlation, which always takes values between -1 and 1, describes the strength of the **linear** relationship between two variables. We denote the correlation by R.



If the linear relationship is strong and positive, R will be near +1.
If it is linear relationship strong and negative, R will be near -1.
If there is no apparent linear relationship, R will be near zero.

Fitting Techniques

- The most simple technique is to fit the data by eye: draw a line in a scatterplot that better describes the relationship
- A more rigorous approach is to make the residuals as small as possible. The more common practice is to choose the line that minimizes the sum of the squared residuals; this technique is called ordinary least squares (OLS).
- This model is referred to as the classical linear regression model (CLR model)
- The parameters estimated (intercept and slopes) using the OLS methodology are considered ‘optimal’

Why OLS?

- It is the most commonly used method
- Any statistical software will be able to apply this technique (including Excel)
- OLS is the benchmark against which all other methods are compared
- It gives a higher weight to points with higher residuals (by squaring the errors)

The first three reasons are historical, traditional and convenient. The last one focuses more on the usefulness of the model.

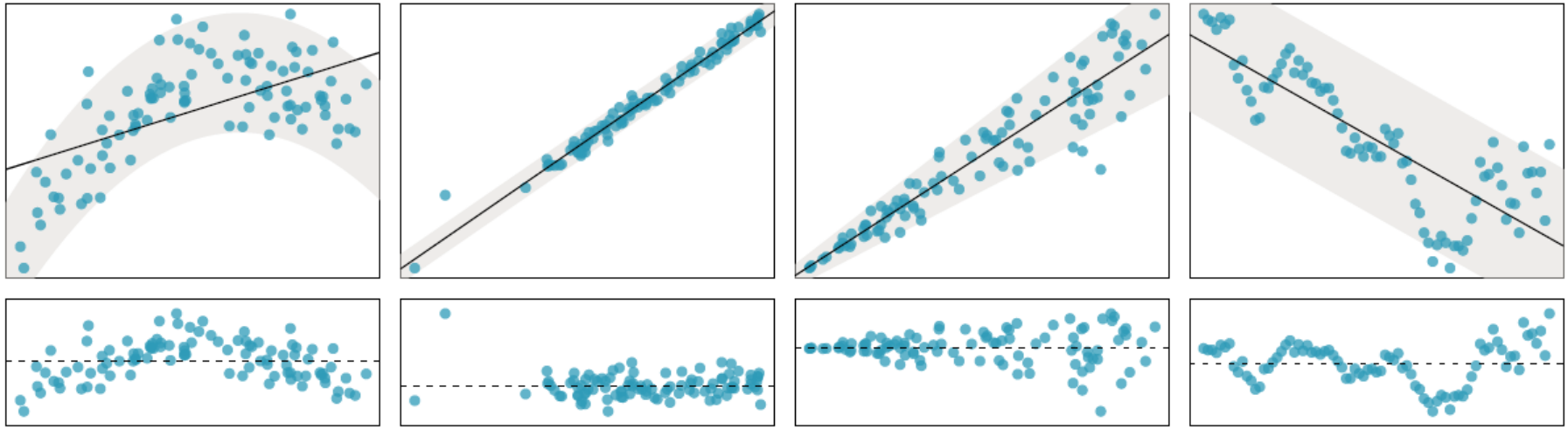
How to Evaluate a Linear Model

The fact the we minimized the square of the residuals does not mean this is the only criterion we use. We also require that:

1. Straight lines should only be used when the data appear to have a linear relationship
   - We will cover non-linear relationships later
2. The residuals are nearly normally distributed & centered on 0.
   - Failure to meet this requirement may indicate a mis-specified model, i.e. it does not include all the important explanatory variables
   - Or there may be influential points
   - If the mean is not zero the model is said to be biased
3. The variability of the residuals is constant along the whole range
4. The observations are independent, i.e. there is no underlying structure in the data

Choosing the Right Model

Below is a graphical representation of different criteria violation.



Relationship not linear	Influential Points (Leverage)	Residual variability not constant	Residuals are auto-correlated

How to Estimate the Parameters of a Linear Model

- In real life applications it is done by using a computer
- A quick estimation can be done by applying:
  - $\beta_1 = \frac{\sigma_y}{\sigma_x} R$ where σ is the standard deviation
  - Knowing that the point $(\bar{x}; \bar{y})$ is on the least squares line, we can calculate β_0 as $\beta_0 = \bar{y} - \beta_1 \bar{x}$

Interpreting Regression Estimates

- The slope describes the estimated difference in the y variable if the explanatory variable x for a case happened to be one unit larger
- The intercept describes the average outcome of y if x = 0
  - This applies only if the model is valid all the way to x = 0, which in many applications is not the case

Take into account the following considerations:

- Model interpretation is the most important step (it is the prize!)
- Data relationship does not mean causality
  - An association or correlation does not necessarily mean a causal connection
- Always interpret the model within the range of your data
  - In technical terms, interpolate never extrapolate

Interpretation of R^2

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line.

- We evaluated the strength of the linear relationship between two variables earlier using the correlation, R. However, it is more common to explain the strength of a linear model using R^2, called R-squared.
- It describes how closely the data cluster around the linear fit

In the book example we will use to practice $\sigma_{aid}^2 = 29.8$ and R^2 = 0.25, 29.8% x 0.25 = 7.45% of the variation in aid received can be explained by the model using the student's family income

Practice 1/8 -Visual Inspection

We will use the data from exercises 7.6 of the book

1. Read in the data
2. Plot
3. Answer 7.6 Husbands and wives, Part I.

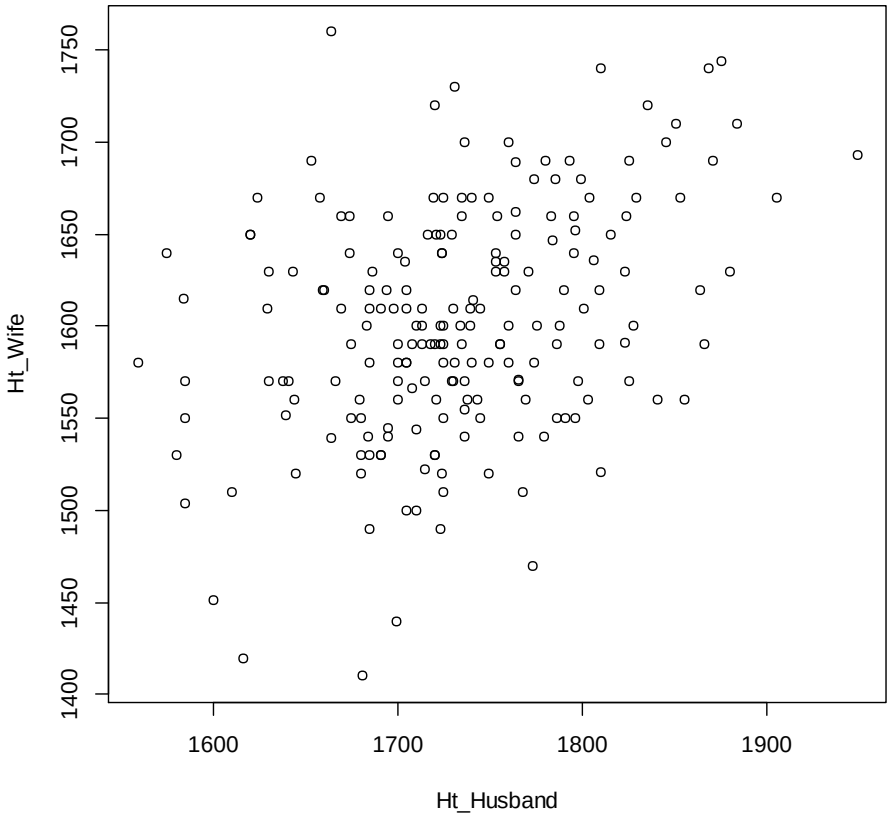
(a) Describe the relationship between husbands' and wives' ages

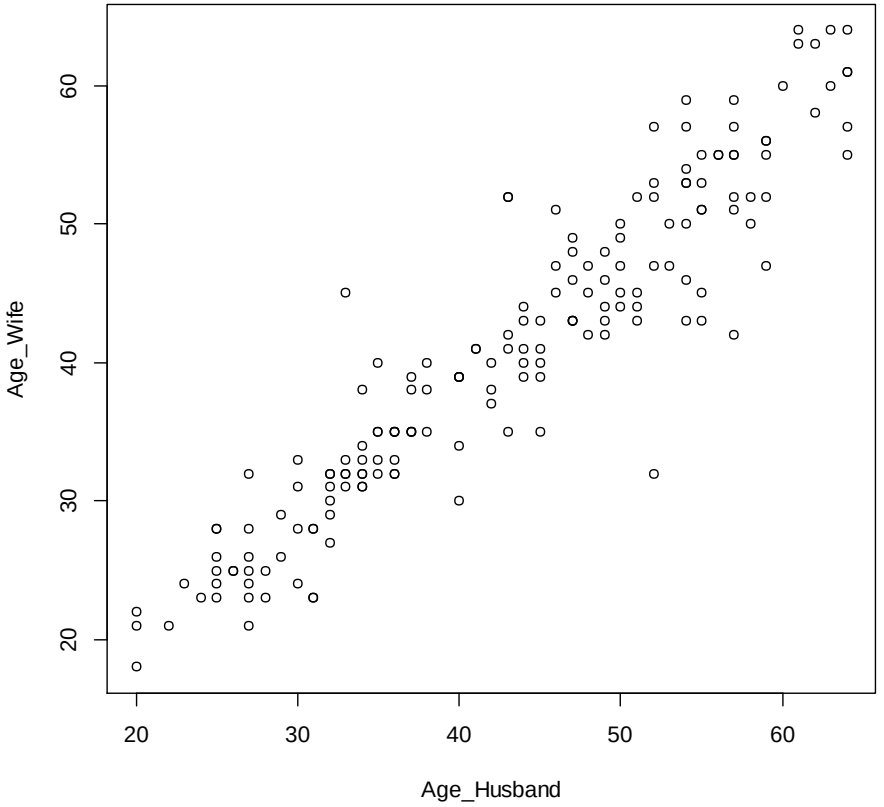
(b) Describe the relationship between husbands' and wives' heights

(c) Which plot shows a stronger correlation? Explain your reasoning.

(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion affect the correlation between husbands' and wives' heights?

Practice 2/8 - Scatter Plot Output





Practice 3/8 - Correlation Example

7.17 Correlation, Part I. What would be the correlation between the ages of husbands and wives if men always married woman who were

(a) 3 years younger than themselves?

(b) 2 years older than themselves?

(c) half as old as themselves?

This can be answered without any calculation, the keyword is ‘always’. Think of an answer before proceeding.

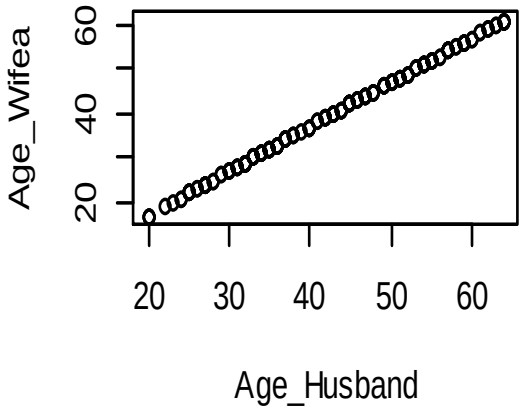
Let’s practice it in R:

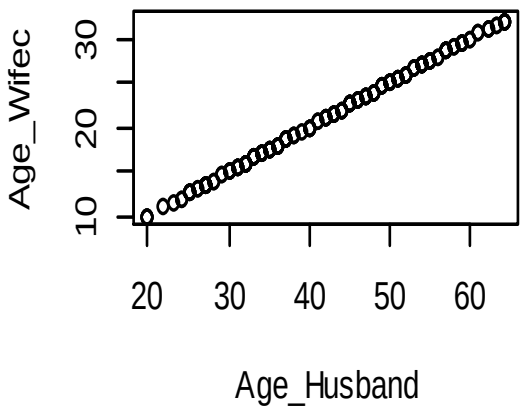
…..

Practice 4/8

First, analyze the data visually

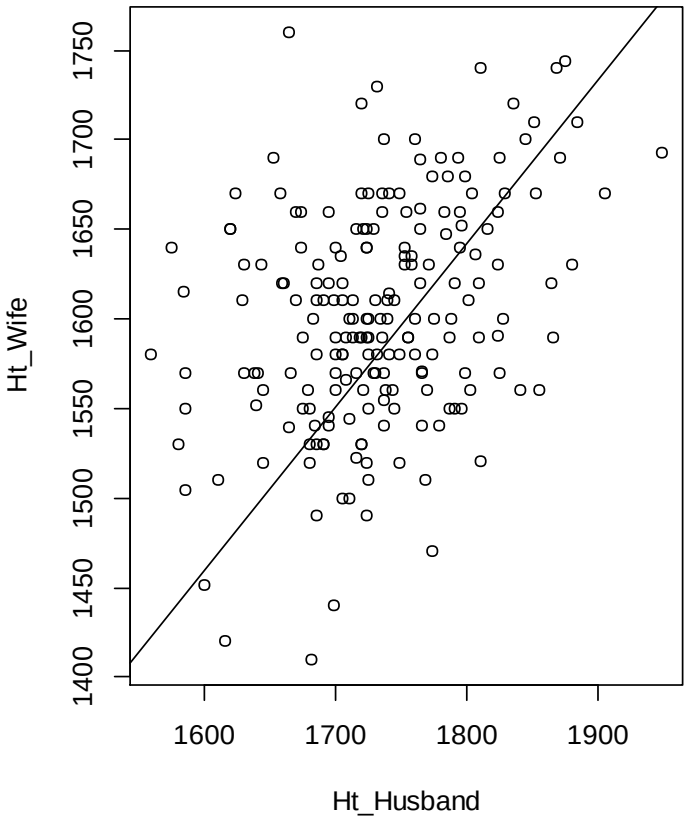
1. Create calculated fields
2. Plot new fields
3. Create scatter plots





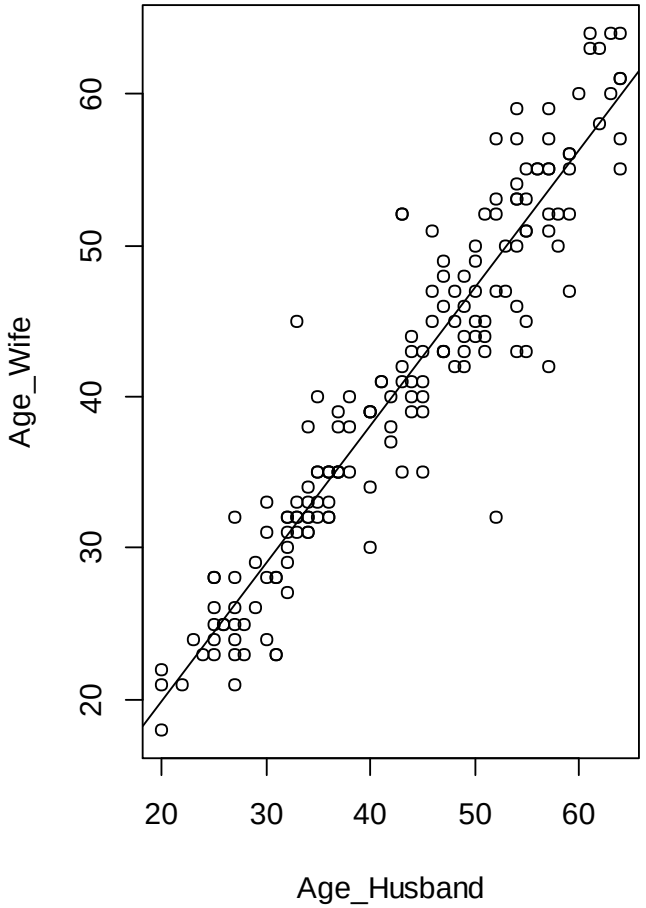
Practice 5/8 - Linear Regression

```
>mHWHt<-lm(Ht_Wife~Ht_Husband)
>summary(mHWHt)

Call:
lm(formula = Ht_Wife ~ Ht_Husband)

Residuals:
    Min      1Q  Median      3Q     Max
-174.91  -40.59   -1.67   42.26  180.72

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.029e+03  1.045e+02   9.846  < 2e-16 ***
Ht_Husband  3.310e-01  6.025e-02   5.493 1.21e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.29 on 197 degrees of freedom
Multiple R-squared:  0.1328,    Adjusted R-squared:  0.1284
F-statistic: 30.17 on 1 and 197 DF,  p-value: 1.212e-07

>par(mfrow=c(1,1))
>plot(Ht_Husband,Ht_Wife)
>abline(lm(Age_Wife ~ Age_Husband))
```



Practice 6/8 - Linear Regression

```
>mHWAge<-lm(Age_Wife~Age_Husband)
>summary(mHWAge)

Call:
lm(formula = Age_Wife ~ Age_Husband)

Residuals:
    Min      1Q  Median      3Q     Max
-16.9586  -1.9897  -0.1035   1.8536  13.3550

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept)   1.57401    1.15012   1.369    0.173
Age_Husband  0.91124    0.02585  35.249   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.951 on 168 degrees of freedom
  (29 observations deleted due to missingness)
Multiple R-squared:  0.8809,    Adjusted R-squared:  0.8802
F-statistic:  1243 on 1 and 168 DF,  p-value: < 2.2e-16
>plot(Age_Wife ~ Age_Husband)
>abline(lm(Age_Wife ~ Age_Husband))
```



Practice 7/8

7.37 Husbands and wives, Part II.

(a) Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test.

(b) Write the equation of the regression line for predicting wife's height from husband's height.

(c) Interpret the slope and intercept in the context of the application.

(d) Given that R2 = 0,09, what is the correlation of heights in this data set?

(e) You meet a married man from Britain who is 5'9" (69 inches). What would you predict his

wife's height to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 6'7" (79 inches). Would it be wise to use the same linear model to predict his wife's height? Why or why not?

Practice 8/8

7.38 Husbands and wives, Part III.

(a) We might wonder, is the age difference between husbands and wives consistent across ages?

If this were the case, then the slope parameter would be 1. Use the information above to evaluate if there is strong evidence that the difference in husband and wife ages differs for different ages.

(b) Write the equation of the regression line for predicting wife's age from husband's age.

(c) Interpret the slope and intercept in context.

(d) Given that R2 = 0.88, what is the correlation of ages in this data set?

(e) You meet a married man from Britain who is 55 years old. What would you predict his wife's age to be? How reliable is this prediction?

(f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife's age? Explain.

Categorical Variables

- There are situations where the independent variables denotes the presence or absence of an attribute.
  - New or old
  - A special event happening like Eastern
  - City were you live
- This situations are considered by using categorical (or dummy) variables.
- The easiest case is a categorical predictor with two levels

```
> marioKart<-read.csv("C:/Users/User/Documents/R/UofT/marioKart.csv",header=T)
> attach(marioKart)
> marioKart
#create dummy variable
> marioKart$cond_num <- as.numeric(marioKart$cond == "new") or > transform(marioKart, condnum = as.numeric(cond == "new"))
> attach(marioKart)
#create scatter plot
> plot(cond_num,totalPr,xlab="condition 0=Used 1=New",ylab="Final Price")
> plot(cond_num,totalPr,ylim=c(30,70),xlab="condition 0=Used 1=New",ylab="Final Price")
```

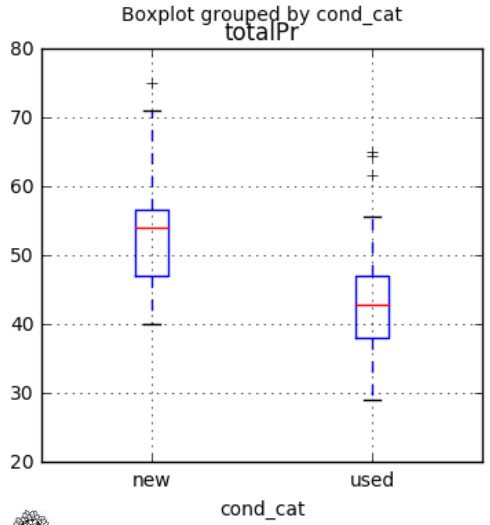
Interpreting Categorical Estimates

The estimated intercept is the value of the response variable for the first category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

- For categorical predictors with just two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance.

Categorical Variables Example

- Let's use MarioKart data from the book and visually analyze if the condition new or used impacts the sales price
- A boxplot plot indicates that there 'on average' the sales price for new is higher than for used
- If we run a regression we see a negative 'slope', i.e. the average price of used is lower than for new



	Coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	53.7707	3.329	16.153	0.000	47.190 60.352
df.cond_cat[T.used]	-6.6226	4.343	-1.525	0.130	-15.209 1.964

Categorical Variables Example

- $53 is the average price for new, while $53-$6 = $47 is the average price for used
- What happens behind the scenes:
  - ‘new’ is the benchmark or reference
  - The slope for ‘used’ indicates the average variation in price when the condition used is present
  - Mathematically, the software creates a dummy variable that assigns 0 to new and to used
  - More of this to come
- We will revisit this topic in the next module, the low R-squared, p value and the interval of confidence shows that the condition is not sufficient to predict the final price (outliers may influence as well)
- In cases where the price is influence by many factors, a multiple regression approach is required – next module

Types of Outliers in Linear Regression

Outliers in regression are observations that fall far from the cloud" of points. These points are especially important because they can have a strong influence on the least squares line.

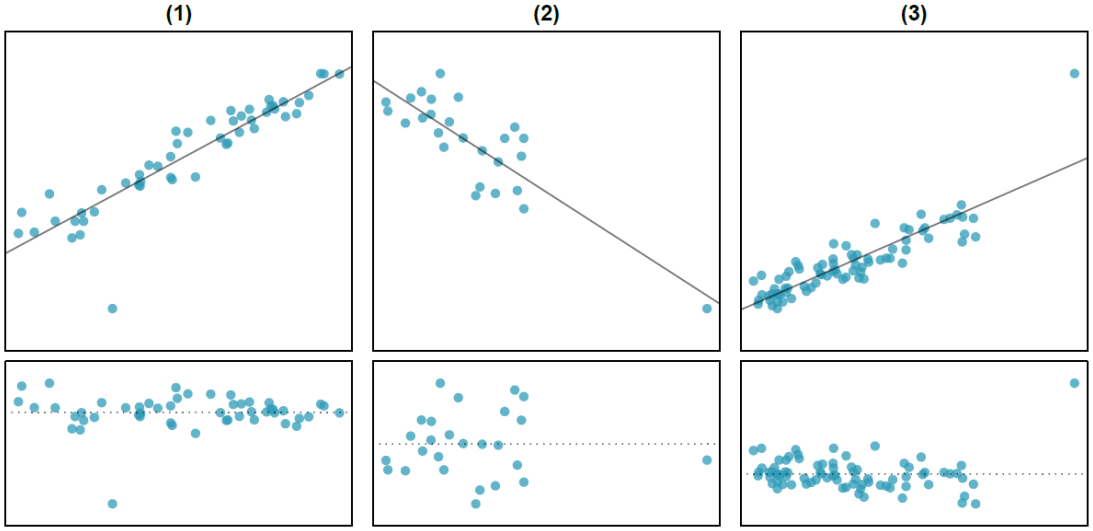
Leverage

Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage**.

If one of these high leverage points does appear to actually invoke its influence on the slope of the line (3, 4, and 5 next slides) then we call it an **influential** point. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

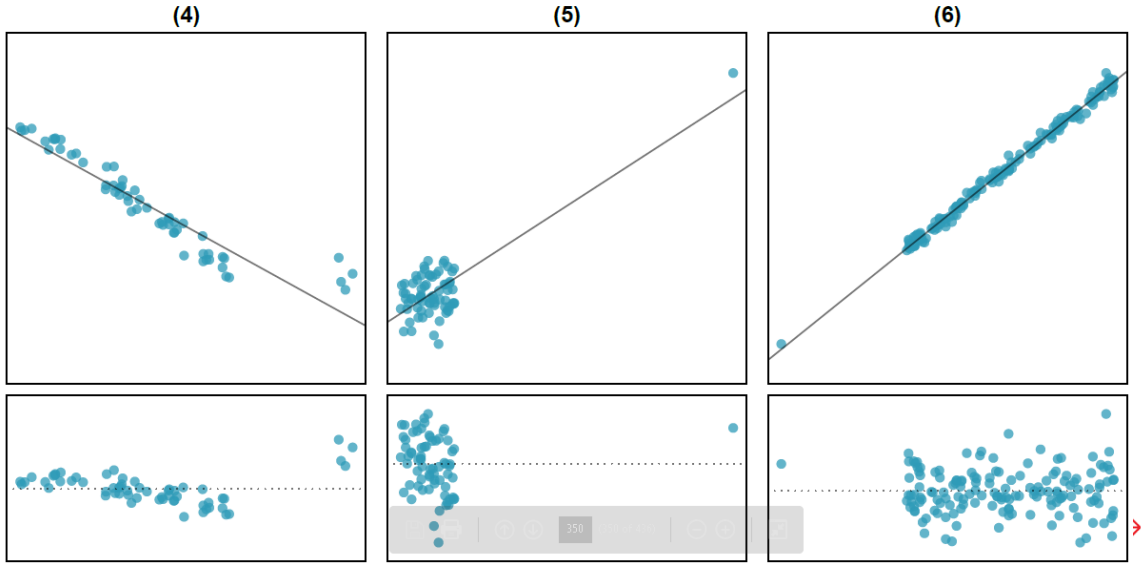
Types of Outliers in Linear Regression

1. There is one outlier far from the other points, though it only appears to slightly influence the line
2. There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential
3. There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well



Types of Outliers in Linear Regression

4. There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere.
5. There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the line
6. There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential



A Word of Caution with Outliers

Caution: Don't ignore outliers when fitting a final model

- If there are outliers in the data, they should not be removed or ignored without a good reason. Whatever final model is fit to the data would not be very helpful if it ignores the most exceptional cases.
- Important information may be contained in the outliers

The objective is not to get a good fit, it is to understand the process you are modeling and to create an useful model. By eliminating outliers you are missing on the opportunity to understand different scenarios you may encounter in the future.

Understanding Regression Output from Software

- Linear equation parameters are estimates, as such we can apply everything we have learnt in hypothesis testing
- The null hypothesis is that the slope are intercept is zero

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.7142	5.4567	-1.23	0.2300
unemp	-1.0010	0.8717	-1.15	0.2617
				$df = 25$

Intercept and slope that minimize the distance between the fitted line and the observations

t Statistics

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-1.0010 - 0}{0.8717} = -1.15$$

p-value
For a two sided test

Caution: If your test is one-sided and the point estimate is in the direction of H_a, then you can halve the software's p-value to get the one-tail area.

Practice

- We will work an example in this and the next 3 modules with simulated data to show a full report crated by Python and its interpretation.
  - A financial institution has income and expenditure data of its clients and wants to perform a risk analysis to help decide to which customers offer a default mortgage insurance

In the Python code the file has been named ‘CreditRisk.csv’

Microsoft Excel
ma Separated Valu

C:\Users\Rodolfo\
ktop\LinearRegres

Description of the Data

- **City:** primary residence of the customer
- **CC Payments:** annualized Credit Card payments or financing the customer did last year. Expenses attributed to categories that are not critical (avoidable)
- **Wage:** annualized wage deposited as direct payments from employees
- **Cost Living:** annual expenses in categories related to groceries, etc… These are considered basic expenses that cannot be avoided
- **Mtg:** Mortgage annual payments
- **Default:** did the customer defaulted his/her mortgage payment last year
- **Vacations:** annual card present expenses done outside of the primary residence. These expenses can be avoided.

Report Interpretation

OLS Regression Results			
Dep. Variable:	df.Wage	R-squared:	0.309
Model:	OLS	Adj. R-squared:	0.307
Method:	Least Squares	F-statistic:	133.4
Date:	Wed, 01 Feb 2017	Prob (F-statistic):	9.44e-26
Time:	19:06:23	Log-Likelihood:	-3187.5
No. Observations:	300	AIC:	6379.
Df Residuals:	298	BIC:	6386.
Df Model:	1		
Covariance Type:	nonrobust		

Report Interpretation – cont.

- **R^2**: describes the amount of variation in the response that is explained by the least squares line or model. It is usually used to describe how well the model fits the data
- **Adjusted R^2** :when you use multiple predictors, as you we will see in the next class, any addition will always improve the R^2 because the degrees of freedom decreases. This indicator balances the number of observations vs. the number of predictors. In practice, be cautious if R^2 and **Adjusted R^2** are very different (we will speak later about overfitting)
- **F Statistics:** like in an hypothesis test, the OLS tests intercept and slope against the null hypothesis, i.e. they are equal to zero. The F-stat is the output of that test.
- **Prob(F-statistic):** the probability of obtaining the slope/intercept produced by the regression assuming the null hypothesis is true. This is also referred as the alpha error.

Report Interpretation – cont.

- **Log-Likelihood:** in short this is the logarithm of sum of the square errors. However, Maximum Likelihood is a whole topic in itself. It is beyond the level of this course but a few tips so if you come across to the term you know what is it about:
  - Likelihood techniques use iterative methods to minimize the square errors of a given model, or the logarithm to help numerical algorithms
  - The postulated model is not constrained by the OLS assumptions
  - It is becoming more popular with the increased computing power available
  - The algorithm start by roughly estimating the parameters, computing predicted values and calculating the sum of squares of the errors, also referred as Likelihood Function
  - Then, by iteratively changing the parameters in small increments, determine the coefficients that minimize the errors
  - The initial estimation is critical to ensure that a global maximum is reached and not a local one

Report Interpretation – cont.

- **AIC and BIC:** Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are based on the log-likelihood function
  - They are used in a similar manner than the Adjusted R^2, to compare models or to judge the usefulness of a newly added variable (avoid over specification)
  - When comparing models, the one with the lowest AIC and BIC should be chosen
- **Covariance type:** beyond the level of this course. Also known as the sandwich method, the robust covariance matrix estimator indicate if certain claims can be done on the covariance matrix

Report Interpretation – cont.

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	6.62e+04	1753.250	37.761	0.000	6.28e+04 - 6.97e+04
df.Cost_Living	0.5318	0.046	11.550	0.000	0.441 - 0.622

The information contained in this table is similar to the report of an hypothesis test:

- Coefficients: slope and intercept that minimize the sum of the square errors, i.e. the estimated values of your model
- Standard error: it is the standard deviation of the errors
- t: value of the t-student distribution of the coefficients
- P>|t|: p value of t being bigger than the calculated value.
- 95.0% Conf. Int.: 95.0% Interval Confidence of the parameters estimated

Report Interpretation – cont.

Omnibus:	13.055	Durbin-Watson:	0.801
Prob(Omnibus):	0.001	Jarque-Bera (JB):	13.041
Skew:	0.471	Prob(JB):	0.00147
Kurtosis:	2.607	Cond. No.	1.16e+05

Some of the following parameters are more advanced. Don't feel bad if it is a bit cumbersome in the beginning.

- **Omnibus Test:** this test uses skewness and kurtosis to test the null hypothesis that the distribution of the errors is normal.
  - A very small value for Pr(Omnibus) indicate a non-normal distribution
- **Jarque-Bera:** another test that considers skewness (S), and kurtosis (K).

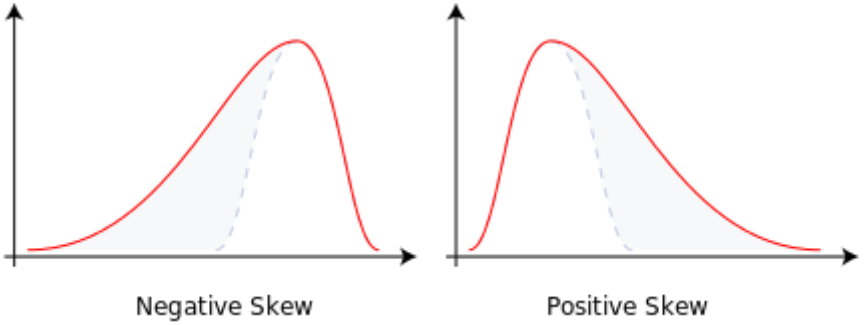
Report Interpretation – cont.

- **Durbin-Watson:** this test performs a mean successive differences test for randomness for a univariate data set looking for autocorrelation
  - If the data are random and from an underlying normal distribution, the average value of the test is 2
  - Values greater than two suggest negative correlation, and values less that one suggest positive correlation
- **Autocorrelation:** is the correlation is between two values of the same variable at times t and t+k. The k value is referred as lag. Notice the difference with linear models, where we want to correlate two different variables
  - the term is widely used in time series, where the values of a random variable is collected at fixed intervals of time
  - In time series the sequence of the data is very important
  - The presence of autocorrelation suggests that a non-linear or time series model be a more appropriate model for these data than a simple constant plus error model

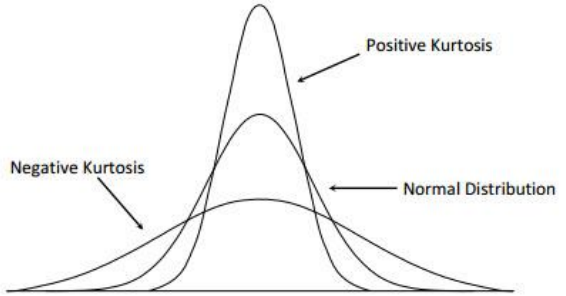
Report Interpretation – cont.

Skewness and Kurtosis: These parameters were described when studying distributions. We will review them to fix concepts

- Average and Variance measure the location and variability of a data set.
- Skewness measures the (lack of) symmetry of a distribution
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.



Graph from:
https://en.wikipedia.org/wiki/Skewness#/media/File:Negative_and_positive_skew_diagrams_(English).svg



Graph from
http://stats.stackexchange.com/questions/84158/how-is-the-kurtosis-of-a-distribution-related-to-the-geometry-of-the-density-fun

Report Interpretation – cont.

- **Condition Number:** the condition number measures the sensitivity of a function's output to its input. Let's explain this a bit:
  - In the next module we will study the case when the dependent variable DP is a function of multiple independent variables IV
  - The IV can be correlated among them – collinearity-, this makes the model extremely unstable, small variation in the inputs can shift the model significantly
  - This parameter evaluates the presence of collinearity
  - A value over 30 would indicate strong collinearity

Sum of Squares

- Variances, not standard deviations are additive, i.e. Var(a+b) = Var(a) + Var(b)
- The reason is that the sum of squares can be partitioned
- In the case of a linear regression: $y_i = \hat{y}_i + e_i$

$SS_{Total} = \sum(y_i - \bar{y})^2 = \sum(\hat{y}_i + e_i - \bar{y})^2 =$

$= \sum(\hat{y}_i - \bar{y} + e_i)^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum e_i^2 =$

$= SS_{Explained} + SS_{Residuals}$

The math is simple and can be found almost everywhere

https://en.wikipedia.org/wiki/Partition_of_sums_of_squares

A Deeper Look to R^2

- R squared is defined as

$R^2 = 1 - \frac{SS_{Residuals}}{SS_{Totals}}$

If $SS_{Residuals} = SS_{Total}$ the model explains nothing

If $SS_{Residuals} = 0 \Rightarrow SS_{Total} = SS_{Explained}$ that is the model can explain all the variability of the dependent variable

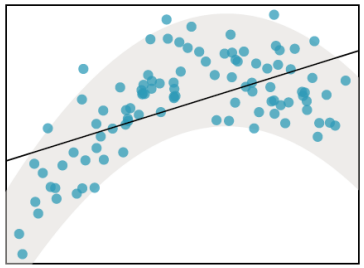
Intercept Removal Effect on R^2 or F

$$R^2 = 1 - \frac{SS_{Residuals}}{SS_{Totals}} = \frac{SS_{Explained}}{SS_{Totals}}$$

- A full model with intercept guarantees that the model is unbiased, i.e. $\sum e_i = 0$
  - If $\sum e_i \neq 0$, you can add its value to the slope and make it zero
- If we force the intercept to be 0, the fitted line will cross the origin when it does not, both $SS_{Explained}$, SS_{Total} (and $SS_{Residual}$) will increase
- $SS_{Explained}$ increases relatively more than $SS_{Residual}$ leading to the increase in R^2 values (see references)
- This illustrates how critical is to always perform a visual analysis

Linear Model Generalization

- Sometimes a non-linear relation can be visualized from the data



- In some cases the data can be transformed so that the transformed data fits a line
- Once you have a good model of the transformed data, the data can be transformed back

Data relation	Equation	Transformation
Exponential	$y = A\ e^{bx}$	log (y) = log (A) + b log (x)
Power	$y = A\ x^2$	log (y) = log (A) + 2 log (x)
Square roots, 1/x and asymptotic 1/(1+x) are special cases of power transformations		

Evaluation of Linear Models

well…, any model

- When the analyst is given a set of data, the common practice is to split the data in two:
  - One set is used to train the model, i.e. determine the coefficients
  - A second set is used to evaluate how efficient is the model to predict data points not present in the training data
- Predictive efficiency simulates the use of the model in real life, i.e. predict unknown dependent variables given possible independent variables
- The actual values predicted should be within the interval of confidence of the predicted value for the model to be useful

Further references

Maximum Likelihood:
http://itl.nist.gov/div898/handbook/apr/section4/apr412.htm
https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

Python example and a good overview of a report:
http://connor-johnson.com/2014/02/18/linear-regression-with-python/

Skewness and Kurtosis:
http://www.itl.nist.gov/div898/handbook//eda/section3/eda35b.htm

Durbin-Watson:
http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/msdt.htm

Autocorrelation:
http://www.itl.nist.gov/div898/handbook//eda/section3/eda35c.htm

Force the intercept to zero:
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/noconstant.htm
http://stats.stackexchange.com/questions/171240/how-can-r2-have-two-different-values-for-the-same-regression-without-an-inte/171250#171250

Exercise

- For each data set in the excel file
- Plot x vs. y and a linear fitted line
- Answer the following:
  - Is a linear model appropriate? Explain.
  - Are outliers present? Yes/No. If there are outliers, do you expect them to be influential? Why?
- Perform a linear regression as y = f(x)
  - Do a graphical analysis (plot histogram of residuals, residual vs. fitted value, y vs. x in the order of collection)
  - Are the 4 assumptions met (slide 4)

Exercise cont.

- If the assumptions are not met (the relation does not look linear
  - Try to perform different transformations (log y, log x)
  - Identify if any transformation makes the transformed data to be linear
  - If so, perform a linear regression
- In all data sets, choose a model and interpret slope and intercept (what happens when x increased by 1?)

Microsoft Excel Worksheet

Next Class

- Multiple Regression
- Logistic Regression
- Complete reading OpenIntro Statistics