

---

# **Quantitative Research Project**

## **Running your own regression project**

Studenmund (2013). Using econometrics: A practical guide (6<sup>th</sup>. edition). Edinburgh: Pearson Education Limited. Chapter 11.

# Stages in the research project

---

- Stage 1: Conduct a **literature review** and formulate an appropriate **research question**.
- Stage 2: Specify a theoretical model about the relationship between the variables (dependent and independent) and formulate **hypotheses/predictions**.
- Stage 3: Collect the **data** (primary or secondary) and prepare them for analysis.
- Stage 4: **Statistical analysis**: Explore the data, conduct linear regression analysis in SPSS, test assumptions, describe the fit of the model and interpret the regression coefficients.
- Stage 5: Formulate a **conclusion and discuss the results**.
- Stage 6: Write a scientific **paper** (max 5000 words not including appendix) about the study you have conducted.

# Choosing a topic

---

- Three key elements when choosing a topic:
  - Try to pick a subject that you find interesting and know something about
  - Make sure that **relevant data are readily available** with a reasonable sample size
  - Look for topics that address an inherently interesting economic or behavioral question or choice.
- On Toledo you will find SPSS datasets + corresponding codebooks for a number of large scale surveys. In addition, many other relevant datasets can be found on the internet.
- As there is limited time, it is recommended that you use available (=secondary) data to conduct a project. E.g.,
  - ESS: <http://www.europeansocialsurvey.org>
  - ISSP: <http://www.issp.org>
  - EVS: <http://www.europeanvaluesstudy.eu>
  - GSS : <http://gss.norc.org>

# Choosing a topic

---

When choosing a topic, you should (taking into account the available data) iterate between the first three stages of the research project:

- Look for **available** datasets on a topic that interests you
- Study the codebook of the survey to see which concepts have been measured in the survey (e.g. altruism, well-being, ...) and which background variables are available (e.g. gender, country, education level,...)
- Conduct a literature review on the concepts that interest you.
- Try to derive a research question that can be answered by applying multiple linear regression on the available data.
- Specify the model:
  - Make sure you have a **quantitative dependent variable**
  - Evaluate whether the available data includes all the **independent variables needed to answer the research question**

# Stage 1: literature review and research question

---

The literature review and research question are needed to write the **introduction of the paper**. A good introduction should include the following elements:

- **Describe the context** of your study. Start broad and then narrow down to make it more specific. Provide some key references.
- **Describe the specific problem statement** on which you will focus in your paper:
  - which aspects have not been studied enough in the literature? What is the gap in the literature?
  - Provide some key references to situate your research in the literature.
  - Describe how your paper will contribute to the existing literature.

# Stage 1: literature review and research question

---

- **Formulate the research question**
  - A sentence that ends with a question mark
  - Make sure the question is specific enough and matches the title of your paper.
- Indicate briefly **how you will try to solve the question**
  - which type of methodology will you use?
- Describe the **added value of your research**
  - Why is it important that you conduct this research? What is the value for the economy? For society?
  - Who will profit from your research?
- End with a short paragraph where you describe **which sections will follow in the rest of the paper.**
- As an illustration we look at the introduction of a paper of Dvorak (2007) on the effect of pay inequality on team performance.

# Example: introduction

Does pay inequality within a team affect performance?

Tomas Dvorak\*

Title: what is the paper about?

## 1. Introduction

The business of sports draws considerable attention from the media and the general public. Fans and sports writers frequently speculate about the effects of money on athletic performance. There is general agreement that more financial resources usually lead to better athletic performance. In team sports, higher pay can be used to lure better players from other teams and therefore improve performance. However, performance can also be affected by pay inequality among players within a team. On the one hand, pay inequality could have a negative effect because it may hinder cooperation among team members. In many sports, team cooperation is critical for good performance. If pay inequality creates tensions or animosity among team members, performance is likely to suffer. On the other hand, inequality could have a positive effect on performance by providing incentives. The prospect of a very large salary could be a powerful drive behind an athlete's performance. Pay inequality might also enhance performance if low paid players learn from high paid players. This would happen when pay inequality is associated with skill inequality. For example, if a highly paid superstar can teach other players, the overall performance of a team may improve. Given that arguments can be made both ways, it is not surprising that there is little agreement on the effects of pay inequality on team performance. The purpose of this paper is to determine whether, on balance, the effect of pay inequality on performance is positive or negative.

Broader view

Research question



# Example: introduction

Understanding the effect of pay inequality on a team's performance is important for at least two reasons. First, team managers can use this information to make decisions about which players to hire. For example, should they hire one expensive superstar and two inexpensive players, or three medium-priced players? If we find that pay inequality leads to poor team performance, then the team may perform better with three medium-priced players than one superstar and two low-priced players. Second, because salaries are a large part of contract negotiations between player associations and team owners, understanding the effects of pay inequality on performance can help determine optimal policies. For example, if pay inequality has a negative effect on performance, an argument for a higher minimum salary could be made.

Importance of research question and added value:

Why is it important to study this topic?

Who will profit from the results of this study?



# Example: introduction

There are a number of studies that look at the effects of pay inequality on performance. DeBrock, Hendricks and Koenker (2004) study the effects of pay inequality on performance in Major League Baseball (MLB). They find that pay inequality is associated with poor performance. Frick, Prinze and Winkleman (2003) look at the effects of pay inequality in all four major leagues in North America. They find that inequality improves team performance in basketball and worsens team performance in baseball. They find no statistically significant effect of inequality on performance in football and hockey.

Results from other studies

This paper looks at the effects of inequality on performance in MLB. It differs from that of DeBrock, Hendricks and Koenker (2004) in that it uses the most recent data. While the previous authors use data from 1985 through 1998, I use data from the latest two seasons: 2003 and 2004. Another difference is that I use a different measure of pay inequality. Rather than the Herfindahl index, I use the percentage of payroll earned by the best paid 20% of players. I chose the share earned by the top 20% players for two reasons: it is somewhat easier to calculate, and its magnitude is easier to interpret.

Situate the paper in the existing literature.

How does the paper contribute to the existing literature?

# Conducting a literature review

---

- Discuss relevant, reliable sources.
- If possible, search for papers published in peer-reviewed journals
- In addition, you can look for reliable internet sources: universities, research centers, research gate, google scholar, ...
- Be critical about sources that are not scientific or less reliable
  - E.g. newspapers, magazines, commercial websites
- Use APA style for citations in the text and in the reference list
  - See APA manual on Toledo
  - Make sure all references cited in the text are included in the list and vice versa.
  - Include complete and correct references in the list in alphabetical order.

# Conducting a literature review

---

- Example references to **journal papers**

Auger, P.B., Devinney, T.M., & Louviere, J.J. (2003). What will consumers pay for social product features? *Journal of Business Ethics*, 42, 281-304.

Burgess, L., Louviere, J. J., & Street, D. J. (2005). Quick and easy choice sets: Constructing optimal and nearly optimal stated choice experiments. *International Journal of Research in Marketing*, 22(4), 459-470.

- Example references to **books**

Lattin, J., Carroll, J.D. & Green, P.E. (2003). *Analyzing multivariate data*. Pacific Grove: Thomson Brooks/Cole.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

# Conducting a literature review

---

Example citations in the text:

- ... the models are described by Lattin, Carroll and Green (2003).
- ... several approaches have been developed to model preference heterogeneity among individuals (Lattin, Carroll & Green, 2003).
- Avoid only summarizing the main findings of relevant papers. Instead, try to **critically integrate** the findings in the existing literature:
  - Identify main trends in the literature
  - Point at important similarities and differences in results of existing studies
  - Compare different studies in a creative way
- Avoid plagiarism
  - Do not literally copy text from other sources, but paraphrase ideas and include a citation to the source.
  - Avoid quotes and definitions: it is not very original.

# Defining the research question

---

The research question should have the following properties:

- It should be clear and specific
- It should include the population on which you will focus in your study (e.g., Flemish consumers)
- It should be economically relevant
- You should be able to answer the question using multiple linear regression analysis
  - E.g. What is the relation between being religious and altruistic behaviour for Europeans?
  - Preferably formulate a research question that can be solved using cross-sectional data (i.e., do not use: How does the crime rate of Brussels evolve over time?)
  - Make sure the necessary data are available and have the correct measurement level (i.e., **quantitative dependent variable**).

## Stage 2 and stage 3: model specification and data

---

Stages 2 and 3 are typically reported in the **Data and Methods section** of the paper. This section typically includes the following elements:

- Describe the source of the data that will be used in the paper
  - Include a correct reference to the data in the text.
  - Discuss the type of sample and give information about the sample size, representativeness etc.
  - Indicate possible shortcomings of the data.
- Describe the variables included in the analysis
  - Describe how each variable is measured/defined. Refer to relevant papers that have included the same type of variable.
  - Describe how the dependent variable is measured or constructed (see scale construction).
  - Use literature, common sense to justify the selection of explanatory variables, do not forget important variables.

## Stage 2 and stage 3: model specification and data

---

- Include **a table that gives an overview of all** variables. The table could include the name of the variable, a description of the variable, the measurement level (quantitative, dummy, categorical) and some descriptive statistics.
  - For quantitative variables include: minimum, maximum, mean, SD, percentage of missing values
  - For dummies or categorical variables include the percentage of observations in each category, the percentage of missing values, which variable is used as the reference when you include the variable using dummy coding



## Stage 2 and stage 3: model specification and data

---

- Discuss what kind of relation you expect between each explanatory variable and the dependent variable.
  - If possible, derive the hypotheses from the literature. Be careful: different theories may predict different signs.
- Discuss how variables will be included in the model
  - For quantitative variables: indicate whether variable transformations may be helpful to reduce skewness or to reduce the influence of outliers.
  - For categorical variables: indicate whether recoding is necessary (e.g., to have enough observations in each category)
- Describe which model(s) will be estimated to answer the research question (e.g., in case you want to conduct a sensitivity analysis to evaluate the impact of different model specifications: different ways to handle missing values, variable transformations, etc.)

# Example: Data and methods section

## 2. Data

The data on pay inequality was constructed in the following way. From the USA Today salary database, I collected salaries for each player in all MLB teams during the 2003 and 2004 seasons. I summed the salaries of all players for each team and each season to obtain the total payroll. The active roster in baseball is 25, but the database includes salaries of disabled players as well. Therefore, the number of players for each team ranges from 25 to 31. As the measure of pay inequality, I calculated the percentage of payroll earned by the highest paid 20% of players. For example, for a 30 player team I summed the salaries of the highest paid 6 players and divide that amount by total payroll. If every player earned the same amount, the best paid 20% would earn exactly 20% of the payroll. When pay is unequal, this measure is higher than 20%. The higher the share of payroll earned by the top 20% of players, the higher the pay inequality.

To measure performance I use the percentage of games won in the regular season. This data comes from BaseballReference.com. It does not include performance during league championships or the World Series. However, with 162 games per regular season, the winning percentage can be regarded as a reasonable measure of performance. This is also the measure used by DeBrock, Hendricks and Koenker (2004).

In addition to pay inequality and performance, I use data on the total payroll of each team. This is a measure of financial resources which could be an important determinant of performance. I measure payroll in current dollars and do not adjust for inflation. While 2003 dollars are not exactly comparable to 2004 dollars, 2003 inflation was low enough not to influence the results significantly.

Where do the data come from ? (In your paper you should use APA style for referencing)

How is each variable defined exactly?

You can refer to other studies to motivate a variable or a technique

This paper investigates the effect of 1 variable (payroll inequality) on performance. The other explanatory variable (total payroll) is added as a control variable.

Indicate possible shortcomings of your data

# Example: Data and methods

---

Table 1 shows the descriptive statistics of each variable. In the first row we see that on average the highest paid 20% of players earn about 61% of the total payroll. This implies that on a 30 player team, the six best paid players earn more than the remaining 24 combined. According to this measure, the team with the most equitable pay is the New York Yankees during the 2003 season when the top 20% of players earned only 42% of total payroll. The team with the highest inequality was the Colorado Rockies during the 2004 season. On that team, five players earned more than 78% of the team's total payroll.

Descriptive statistics are briefly discussed in the text

The second row in Table 1 shows that the average winning percentage is 50% which has to be the case since for every game won there is a game lost. The Detroit Tigers have the lowest winning percentage in the data with only 26% of games won during the 2003 season. The maximum winning percentage in the data is for the St. Louis Cardinals, who won nearly 65% of their games during the 2004 season. Finally, the last row in Table 1 shows that the average payroll is about 70 million dollars. The range of payroll is quite striking. It goes from less than 20 million dollars for the Tampa Bay Rays to over 184 million for the New York Yankees.

Table 1: Descriptive Statistics

	mean	median	st.dev.	min	max
Top20share (in %)	61.0	61.4	8.0	42.2	78.3
Games Won (in %)	50.0	51.6	8.2	26.5	64.8
Payroll (in mil. USD)	70.0	65.3	30.3	19.6	184.2

First glance at the data using descriptive statistics: summarizing numbers



# Extracting a subset of cases/variables from a database

---

- When using data from a large-scale survey you should first extract a small data set that includes the population and the variables you need.
- Selection of population of interest
  - If the database contains data of many countries, it may be unrealistic to assume that the same regression model holds for all countries. As modelling the differences between all countries may be difficult, it is recommended to select respondents of one, or a few similar countries.
  - To select a subset of respondents in SPSS: specify the selection condition in the following menu  
**Data/select cases/if condition is satisfied**
  - To save a new data set that contains only selected cases: run the following statement in a syntax window  
**Save outfile='c:\selection.sav' /unselected=delete.**
- Selection of variables of interest in SPSS  
**File/ save as/variables**

# Model Specification

---

- As you need to use linear regression, the **response variable should be quantitative**. i.e., it should have **enough values and variation**: preferably more than 20 values.
- E.g. income in Euro
  - Do not use [0,1000), [1000,3000), [3000,6000),  $\geq 6000$  (not linear, not enough variation)
  - One could measure income using bins [0,500), [500,1000), ..., [5500,6000]. The bins would be presented by the values 250, 750, ..., 5750.
- In surveys, concepts are often measured using a set of likert-scale items. You can measure the concept by **constructing a scale**, i.e. by summing the items that measure the same concept.
- Scale construction can be interesting, especially to measure the dependent variable.

# Scale construction

---

- Use Cronbach's alpha to evaluate whether the sum of the items is a **reliable measure of the concept**.

- In SPSS: **analyse/scale/reliability analysis**
- Cronbach's alpha is a coefficient between 0 and 1 with higher values indicating that the sum of the items is a more reliable measure of the concept:

$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

# Scale construction

---

- When computing Cronbach's alpha, **make sure that all items are oriented in the same way**. E.g. the following two likert-scale items (1= strongly disagree, .., 7=strongly agree) have a different orientation and hence one of the items should be reversed (replace 1 by 7, replace 2 by 7, etc. ) when computing alpha.
  - The balance of nature is strong enough to cope with the impacts of modern industrial nations.
  - The balance of nature is very delicate and easily upset.
- As an example consider 15 items that measure to what extent a certain behaviour is justifiable. After inspecting the items, we suspect that:
  - The items 1-5 measure to what extent a person would find it justifiable to steal from the government
  - The items 6-12 measure to what extent a person is in favour of rights on self-determination
  - The items 13-15 measure to what extent a person would find violence justifiable



# Scale construction

---

variable	type	label
1 justifiable_claiming_benefits	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Claiming government benefits to which you are not entitled
2 justifiable_avoiding_fare	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Avoiding a fare on public transport
3 justifiable_stealing	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Stealing property
4 justifiable_cheating_taxes	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Cheating on taxes if you have a chance
5 justifiable_accept_bribe	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Someone accepting a bribe in the course of their duties
6 justifiable_homosexuality	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Homosexuality
7 justifiable_prostitution	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Prostitution
8 justifiable_abortion	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Abortion
9 justifiable_divorce	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Divorce
10 justifiable_premarital_sex	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Sex before marriage
11 justifiable_suicide	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Suicide
12 justifiable_euthanasia	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Euthanasia
13 justifiable_beat_wife	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: For a man to beat his wife
14 justifiable_beat_children	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Parents beating children
15 justifiable_violence	10-point likert scale (1=never justifiable, 10= always justifiable)	Justifiable: Violence against other people

# Scale construction

---

- We use SPSS to compute Cronbach's alpha on items that are assumed to measure the same concept. This yields the following result:

items	Cronbach's alpha
1-5	.81
6-12	.89
13-15	.76

- We see that the first two scales have a good reliability, and that the third scale has an acceptable reliability. Hence, for each scale we can compute the sum of the items and use this new variable to measure the concept.

## Stage 4: Statistical analysis

---

The statistical analysis is usually presented in the **results section** of the paper. This section should contain the following elements:

- Describe the **univariate relation between each independent variable and the dependent variable**.
  - For **quantitative predictors** you can compute a **sample Pearson correlation** between the predictor and the dependent variable and test whether the correlation is significantly smaller than/larger than/different from 0.
  - As an alternative, you could use **simple regression** and look at  $R^2$  to study the strength of the linear relation between a predictor and the dependent variable.
  - Remark: make a scatterplot of the relation between the variables, or a residual plot to check whether the relation is linear, to check for important outliers, etc.

# Statistical analysis

---

- To study the relation between a **binary independent variable** (e.g. gender) and the dependent variable you can use an **independent samples t-test** (e.g., you test whether the average scores of males and females on the dependent variable differ significantly).
- To study the relation between a **categorical independent variable** and the dependent variable, you can use **one-way anova**.
  - You can use an F-test to test whether the average score on the dependent variable differs across the levels of the categorical independent variable.
  - You can report the effect size measure eta-squared to indicate how much of the variation in the dependent variable is explained by the categorical independent variable.
- As an alternative, you can **run multiple linear regression** on the dummy coded categorical independent variable and look at  $R^2$  to summarize the strength of the relation.

# Statistical analysis

---

- Use **multiple linear regression** to investigate how each independent variable affects the dependent variable after controlling for the other variables in the model. For each model, the following information should be included:
  - Estimated regression coefficients, standard errors, t-values and p-values
    - Round off estimates to an appropriate number of digits
    - use an appropriate scale for each variable (e.g., in 1000 units).
  - Sample size data, unit of analysis data, unit of measurement for each variable, coding used for categorical variables
  - $R^2$ ,  $R^2_{adj}$ , residual standard error
- Remark: do not just paste tables of SPSS in the paper, but present the results in an overview table with layout in APA style.

# Report results of the regression analysis: Example

---

Regression of the Sales of a product (in 1000 units produced) in 200 markets as a function of advertisement budget spent in different media (in \$1000), namely TV, Newspaper and Radio.

Dependent variable: Sales (in 1000 units produced)

Variable	$\hat{\beta}$	$SE(\hat{\beta})$	t	p-value
constant	2,939***	0,312	9,42	0,000
TV budget (in \$1000)	0,046***	0,001	32,81	0,000
Newspaper budget (in 1000\$)	0,189***	0,009	21,89	0,000
Radio budget (in 1000\$)	-0,001	0,006	-0,18	0,860

$$R^2 = .897, R_{adj}^2 = .896$$

$$N = 200$$

$$*p < .05, ** p < .01, *** p < .001$$

# Statistical analysis

---

## Check the assumptions of the regression model(s).

- Use a residual plot (unstandardized predicted values against standardized residuals)
  - to evaluate the **linearity assumption**
  - to check for **heteroscedasticity**, if doubtful conduct the test of White.
- In case there is nonlinearity, or heteroscedasticity: Use appropriate variable transformations or model extensions to obtain a valid model.
- If the data set is small: use a Normality test to evaluate whether the residuals are **Normally distributed**.
- Remark: if you have to make several modifications to obtain a valid model, discuss the steps taken in the paper, and include residuals plots of the final model. You might include other information in the appendix.
- Check for multicollinearity: include VIF statistics



# Statistical analysis

---

- Conduct a **sensitivity analysis** to compare different model specifications
- This may be interesting if you are in doubt about the specification of the functional form of certain variables (log transformation or not), how to deal with missing values (e.g., replace by mean or listwise deletion) etc.
- If you want to investigate whether removing a non-significant variable is problematic.
- Discuss the results of the final model
  - Include a technically correct interpretation of the (significant) estimated regression coefficients
  - include information on practical significance of the predictor: e.g. check how much  $R^2$  would decrease if the predictor is dropped from the model.
- As an illustration we look at the results section of the paper of Dvorak (2007).

# Example: results

## 3. Empirical Results

I estimate three different specifications. The dependent variable in each specification is performance, as measured by the percentage of games won. Pay inequality and total payroll are the independent variables. Table 2 shows the results. In the first specification, I regress performance on the share earned by the top 20% of players. The coefficient on the share of top 20% is negative and statistically significant. This indicates that teams with higher pay inequality tend to win fewer games. A one percentage point increase in the share of payroll earned by the top 20% of players is associated with about half of a percentage point decline in the percentage of games won.

Discuss size and sign of the coefficients, if significant

Table 2: Regression Results

Dependent variable: winning percentage (in %)			
	(1)	(2)	(3)
Intercept	77.3 (7.43)**	59.9 (9.51)**	37.35 (15.37)*
Top20share (in %)	-0.45 (0.12)**	-0.27 (0.13)*	-0.28 (0.13)*
Payroll (in mil. USD)		0.10 (0.04)**	
Log of Payroll			7.09 (2.43)**
R-squared	0.19	0.29	0.29
Adjusted R-squared	0.18	0.26	0.27

Number of observations is 60.

Standard errors are in parentheses.

\*\* significant at 1%, \* significant at 5%

Results of a regression are usually presented in a table like this containing:

- For each explanatory variable: estimated coefficient, t-value or standard error, significance here using \*
- Also the global goodness of fit is presented using  $R^2$ , (adjusted  $R^2$ )

## Example: Results

In the second specification I include total payroll as an independent variable. Payroll is a measure of the financial resources which can affect performance - the higher the payroll, the higher the quality of players and, generally, the better the performance. Therefore, including payroll may increase the precision of the estimated coefficient on pay inequality. More importantly, it is possible that pay inequality is correlated with total payroll. If low payroll teams tend to have more pay inequality, then the coefficient on pay inequality in specification (1) is biased. Indeed, the correlation coefficient between the share earned by the top 20% of players and total payroll is -0.5. Teams with high pay inequality may perform worse not because of pay inequality, but because they are also the teams with a lower payroll. Therefore, in order to measure the effect of pay inequality on performance, I need to control for total payroll.

Once I control for total payroll, the coefficient on the share of the top 20% remains statistically significant but the magnitude drops substantially. Holding payroll constant, a one percentage point increase in the share earned by the highest paid 20% is associated with a 0.27 percentage point decline in the percentage of games won. The impact of inequality on performance does not seem enormous. For example, a five percentage point increase in inequality for the team with median inequality would shift the team up 13 spots in the inequality ranking, but its performance ranking would drop by only 2 spots. The coefficient on total payroll is positive and statistically significant. A one million dollar increase in total payroll is associated with about 0.1 percentage point increase in the percentage of games won. This indicates that greater financial resources tend to improve performance. Adding payroll as an independent variable led to an increase in R-squared from about 0.19 to 0.29.

Second model specification: using control variable

Compare model specification: adjusted  $R^2$ , Sign, size of coefficients

Not only the statistical significance, but also the economic 'significance' is discussed here



## Example: Results

Finally, in specification (3) I include the logarithm of payroll instead of payroll. I want to verify that the result in specification (2) is robust to different functional forms. In addition, the effect of an additional one million dollars may be smaller for a team with a 100 million payroll than for one with a 20 million payroll. Thus, including payroll in logarithm seems appropriate. The coefficient on the share of the top 20% remains statistically significant with roughly the same magnitude. The log of payroll is statistically significant. A one percent increase in payroll is associated with about 0.8 percentage points increase in the percentage of games won.

Table 2: Regression Results

Dependent variable: winning percentage (in %)			
	(1)	(2)	(3)
Intercept	77.3 (7.43)**	59.9 (9.51)**	37.35 (15.37)*
Top20share (in %)	-0.45 (0.12)**	-0.27 (0.13)*	-0.28 (0.13)*
Payroll (in mil. USD)		0.10 (0.04)**	
Log of Payroll			7.09 (2.43)**
R-squared	0.19	0.29	0.29
Adjusted R-squared	0.18	0.26	0.27

Number of observations is 60.

Standard errors are in parentheses.

\*\* significant at 1%, \* significant at 5%

Third model specification: logarithm of 'total payroll':

- To check robustness of the estimations
- Argue why logarithm transformation of payroll is useful

Revisit the results: comparison of three model specifications. Note that if an explanatory variable is not in the model, the cell is empty.

In this paper the model assumptions are not specifically checked. You should do this in your paper !

## Stage 5: Formulate a conclusion and discuss results

---

The **discussion section** of the paper is used to present conclusions and discuss the results. The section should include the following elements:

- A **summary** of the results. The reader should be able to read this summary without reading the entire paper.
- A discussion of the results:
  - Formulate a clear **answer to the research question**.
  - compare results of the estimated regression model(s) against the **hypotheses** (e.g. about sign of coefficients) stated in the data and methods section.
  - **Compare your results with existing findings in the literature**.
- Discuss the **consequences of the results** (e.g., for stakeholders).
- Discuss the **limitations** and drawbacks of your study.
- Make suggestions for **future research**.
- As an illustration we look at the conclusion of Dvorak (2007).

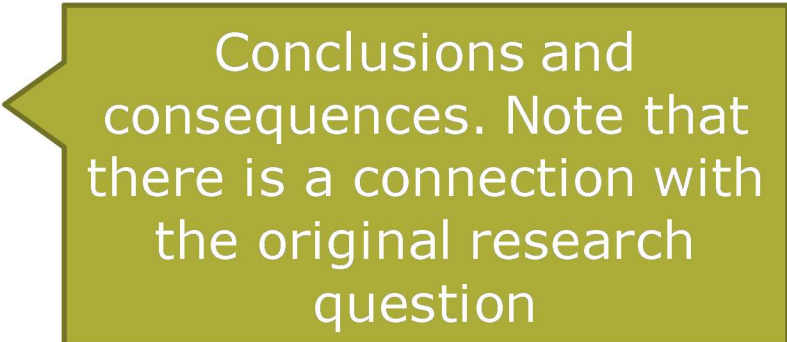
# Example: Conclusion and discussion

---

## 4. Conclusion

The analysis in this paper shows that pay inequality within MLB teams has a negative effect on performance. The effect remains statistically significant even after controlling for total payroll. The result is the same as that of DeBrock et al. (2004) who use data from 1985 through 1998. My paper confirms their finding using the most recent data and using a different measure of pay inequality.

The fact that pay inequality leads to worse performance implies that managers should strive for pay equality in their teams. For example, instead of hiring two low-priced players and one superstar, performance may be better if three medium-priced players are hired. Given these results, it is surprising that there is not a more equal distribution of pay in baseball. One possible explanation is that managers may care about attendance as well



Conclusions and consequences. Note that there is a connection with the original research question



# Example: Conclusion and discussion

---

The conclusions above are subject to a number of limitations. First, it is unclear to what extent the results can be generalized to other sports. Each sport requires a different degree of cooperation among team members. Therefore, the relationship between pay inequality and performance is likely to differ across sports. Second, the error terms for each team could be correlated over time. For example, if a team wins a lot of games one year given its payroll and pay inequality, that team is likely to win a lot of games the next year as well. Therefore, the estimation procedure may need to correct for this autocorrelation. Finally, there may be other variables that affect performance, e.g. coach salary or quality of training facilities. Including these in the regression would increase the precision of my estimates as well as eliminate potential omitted variable bias.

The channels through which pay inequality affects performance are not clear. I can think of two possibilities. One is that pay inequality leads to tensions within the team and impairs performance. The other possibility is that baseball requires players of similar quality. Pay inequality is probably associated with skill inequality, and it may be the skill inequality that drives down performance. An excellent pitcher cannot win the game when the outfielders cannot catch or throw. It may be possible to distinguish these two channels empirically. Using statistics on individual player skill level, one could construct a measure of skill inequality for a team and include it as an additional control. The coefficient on pay inequality in that case would capture the effect of pay inequality on performance while holding skill inequality constant. A negative impact of pay inequality would then support the idea that pay inequality leads to tensions which affect performance. This investigation, however, is left for future research.

## Limitations in your research:

- External: Can you generalize the results to other sports ?
- Internally: shortcomings about the research methodology ?

## Reflection about the research and suggestions for further research



# Example: Conclusion and discussion

---

## References:

DeBrock, Lawrence, Wallace Hendricks, and Roger Koenker. 2004. Pay and performance: The impact of salary distribution on firm-level outcomes in baseball. *Journal of Sports Economics* 5 (August): 243–261.

Frick, Bernd, Joachim Prinz, and Karina Winkelmann. 2003. Pay inequalities and team performance: Empirical evidence from the North American major leagues. *International Journal of Manpower* 24: 472-491.

## Appendix:

Data with documentation and results: [MLB.xls](#)

You should use APA style for referencing.

In this example paper a different style is used.

## Stage 6: writing a scientific paper

---

- The paper should contain at most 5000 words, appendix not included. It includes the sections Introduction, Data and methods, Results, Conclusion and discussion, references, appendix.
- **Layout and style**
  - Use the available template (template\_article.docx)
  - Use APA style for layout of tables, for citations in the text and in the reference list.
  - Give each figure or table a number and a caption (**in MS word: references/insert caption**), and refer to each table or figure in the text using cross reference (**in MS Word: references/cross reference**). Make sure the information in the table can be understood by the reader.
  - Do not paste tables of SPSS in the text, but create appropriate tables using APA layout.
  - You can paste SPSS output in the appendix, but make sure the information is well-structured so that the reader can understand it.

# Stage 6: writing a scientific paper

---

- **Structure of the text:**

- Logical structure: this is not necessarily the order in which you carried out the research.
- Know your reader's level/knowledge.
- Add connecting texts, to make sure that the reader understands the 'flow' of the text.
- Introduce a new (sub)section. Make connections.
- Add appropriate words to structure the text (in addition, furthermore, so that, however etc.)

- **Paragraph structure:**

- A paragraph is a meaningful whole
- Use key sentences at the beginning of each paragraph.
- Do not make them too long/too short (on average 7 lines)

# Stage 6: writing a scientific paper

---

## •Titles

- Be specific: not too generic e.g. ‘Results’, but ‘Results about...’.
- Titles cannot replace text. You should be able to read the text without reading the titles.
- Table of contents: keep the structure simple, do not include many subsections

## • Narrative viewpoint

- Avoid I/we as much as possible
- Do not replace I/we by ‘one’ and avoid passive sentences.
- What ‘players’ are playing the lead in your text? Use these as subject of your sentence: e.g. the employer ...
- Do not directly address your reader!

# Stage 6: writing a scientific paper

---

- **Be specific**

- Use a neutral objective style
- Avoid vague words like ‘a few’, ‘some things’,...
- Avoid IKEA sentences: sentence that you can cut and paste in any paper.  
Do not write: “We will start with a literature review. Based on this we create the theoretical framework for our research. We end with an empirical analysis”.
- Provide the reader with enough examples.

- **Simplicity is a virtue**

- Do not use difficult words unless this is necessary.
- Be concise and to the point: Avoid long opening parts like “We would like to point out that”.. “ in this text we would like to provide you with an answer on ...” Do not announce that you are about to state something. Immediately give the message.

# Stage 6: writing a scientific paper

---

- **Vocabulary**

- Careful choice of words
  - Avoid spoken language!
  - Use a varying vocabulary
  - Spelling: avoid typing errors
- Pay attention to writing sentences that are grammatically correct.
  - Avoid including footnotes, instead integrate ideas and relevant information in the text.