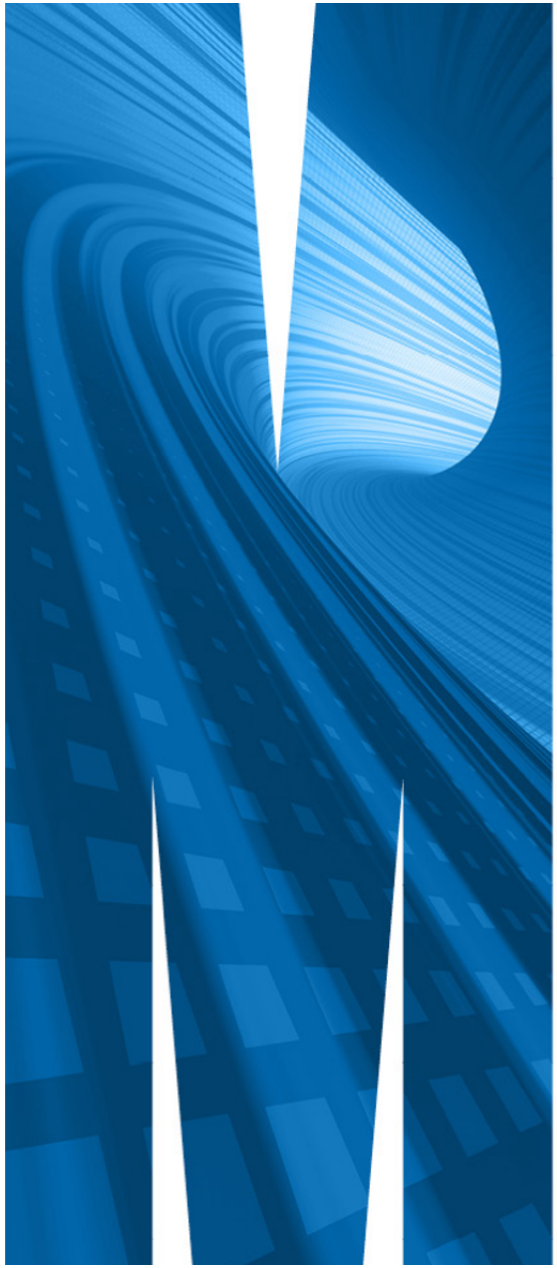
MONASH University

# ETC3250/5250: Introduction to Machine Learning

## Review of regression




Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu


Week 2a

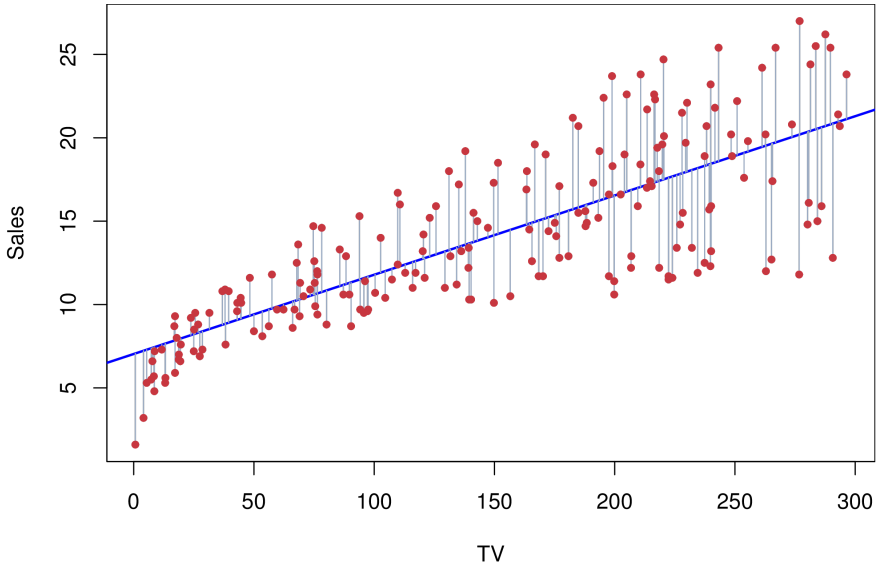
# Multiple Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \ldots, n$$

- Each $X_j$ is called a predictor, or independent variable or input variable.
- The coefficients $\beta_1, \ldots, \beta_p$ measure the effect of each predictor after taking account of the effect of all other predictors in the model.
- Predictors may be transforms of other predictors. e.g., $x_2 = x_1^2$. Then the model form would be nonlinear. Categorical predictors need to be converted into dummy variables (see a few slides along).
- Once we estimate the model, we will have estimated coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$, predicted values, $\hat{Y}$, and residuals, $e_i, i = 1, \ldots, n$.

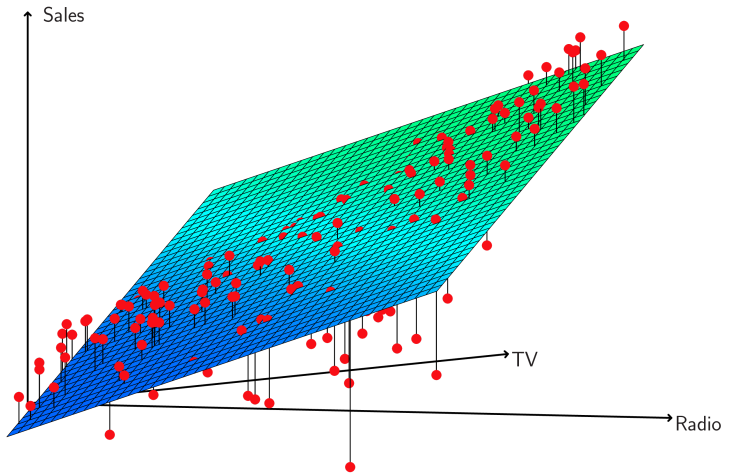
# Multiple Regression

The model describes a **line**, plane or hyperplane in the predictor space.




(Chapter3/3.1.pdf)

The model describes a line, **plane** or hyperplane in the predictor space.




(Chapter3/3.5.pdf)

# Categorical Variables

Qualitative variables need to be converted to numeric, by making a set of dummy variables.

$$x_i = \left\{ \begin{array}{ll} 1 & \text{if } \ i^{\text{th}} \text{ obs is a koala} \\ 0 & \text{otherwise} \end{array} \right\}$$

which would result in the model

$$\hat{y}_i = \left\{ \begin{array}{ll} \beta_0 + \beta_1 & \text{if } \ i^{\text{th}} \text{ obs is a koala} \\ \beta_0 & \text{otherwise} \end{array} \right\}$$

# Categorical Variables

More than two categories

$$x_{i1} = \left\{ \begin{array}{ll} 1 & \text{if } \; i^{\text{th}} \text{ obs is a koala} \\ 0 & \text{otherwise} \end{array} \right\}$$

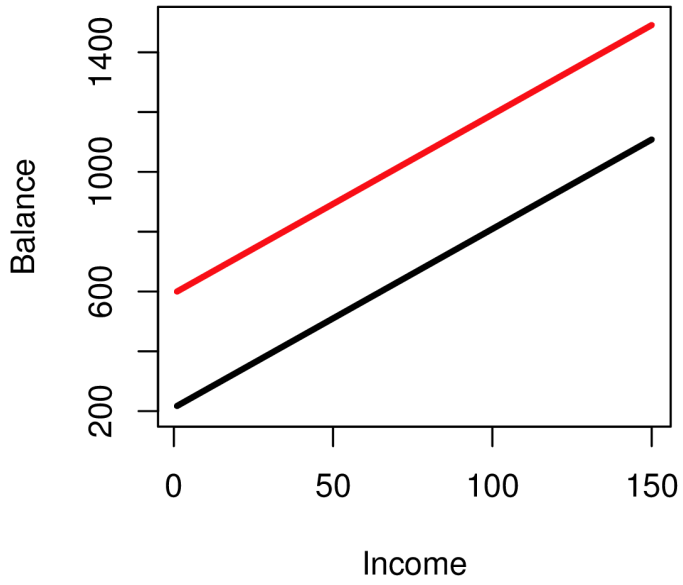
$$x_{i2} = \left\{ \begin{array}{ll} 1 & \text{if } \; i^{\text{th}} \text{ obs is a bilby} \\ 0 & \text{otherwise} \end{array} \right\}$$

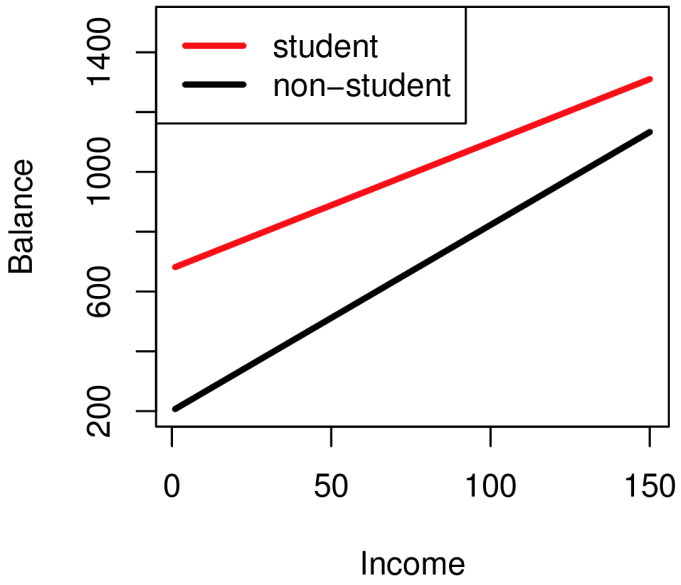
which would result in the model using dummy variables.

$$\hat{y}_i = \left\{ \begin{array}{ll} \beta_0 + \beta_1 & \text{if } \; i^{\text{th}} \text{ obs is a koala} \\ \beta_0 + \beta_2 & \text{if } \; i^{\text{th}} \text{ obs is a bilby} \\ \beta_0 & \text{otherwise} \end{array} \right\}$$

# Interactions are induced by categorical predictors

When you have a categorical variable, it can be convenient to allow BOTH slope and intercept to vary across category levels. This is called an interaction.



# Inference

- Is at least one of the predictors useful in predicting the response?
- Do all the predictors help to explain $Y$, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict and how accurate is our prediction?

## Model fitting

Least squares is a common way to fit the model, where $\hat{\beta}_j$ are chosen to minimise

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

The smaller the sum of differences, the better the model fits the data.

## Goodness-of-fit

$R^2$ is the proportion of variation explained by the model, and measures the goodness of the fit, close to 1 the model explains most of the variability in $Y$, close to 0 it explains very little.

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

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

## Model Diagnostics

Residual Standard Error (RSE) is an estimate of the standard deviation of $\varepsilon$. This is meaningful with the assumption that $\varepsilon \sim N(0, \sigma^2)$.

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

This is another way to examine the variation around the model. Unlike $R^2$ it is not on a standard scale.

# Maximum Likelihood Estimation and Least Squares

If the errors are iid and normally distributed, then

$$Y \sim N(X\beta, \sigma^2 I)$$

So the likelihood is

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \right)$$

which is maximized when $\sum_{i=1}^{n} (Y - \hat{Y})^2$ is minimized.

## Significance tests

An $F$-test can be computed to assess if **any** predictor explains response, by testing this hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \text{ vs } H_a : \text{at least one is not } 0$$

where the test statistic is $F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$.

## Individual variables

The strength of relationship between the response and an individual variable can be tested using a $t$-test for the hypothesis:

$$H_0 : \beta_j = 0 \text{ vs } H_a : \beta_j \neq 0$$

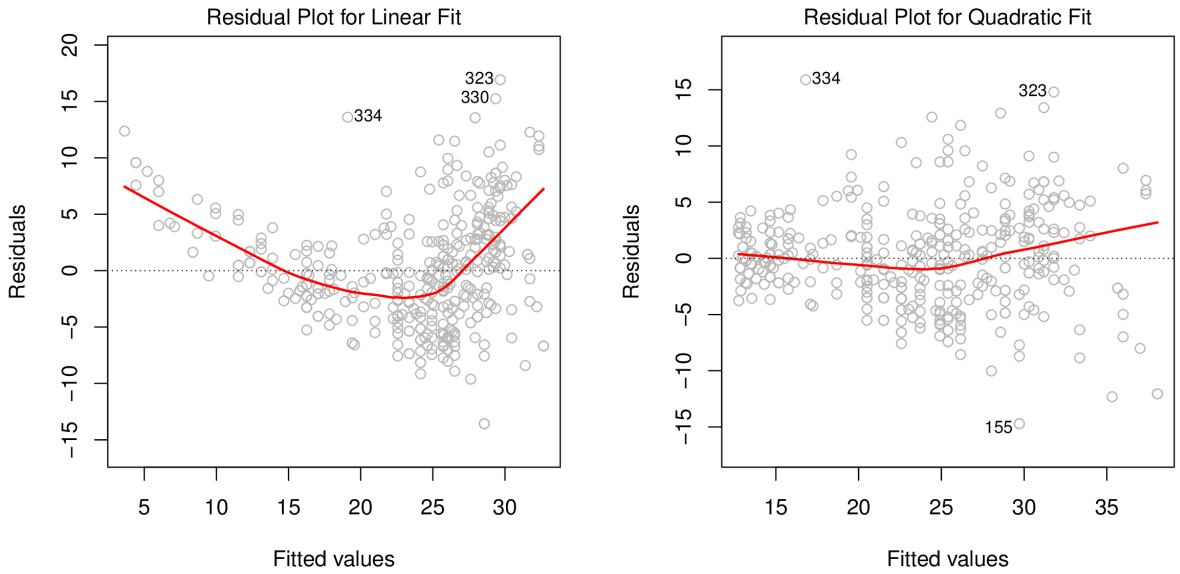
where the test statistic is $t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$

# Interpreting the effect of any predictor

We interpret $\beta_j$ as the average effect on $Y$ of a one unit increase in $X_j$, holding all other predictors fixed.

# Assessing model fit using residuals

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Try a transformation, but may not fix.)




(Chapter3/3.9.pdf)

# Comparing models

An $F$-test can be computed to assess if **any additional** predictors significantly help to explain response, by testing this hypothesis

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0 \text{ vs } H_a : \text{at least one is not } 0$$

where the test statistic is $F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$.

(We are considering a model with just $q$ variables as opposed to all $p$.)

# Common pitfalls

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

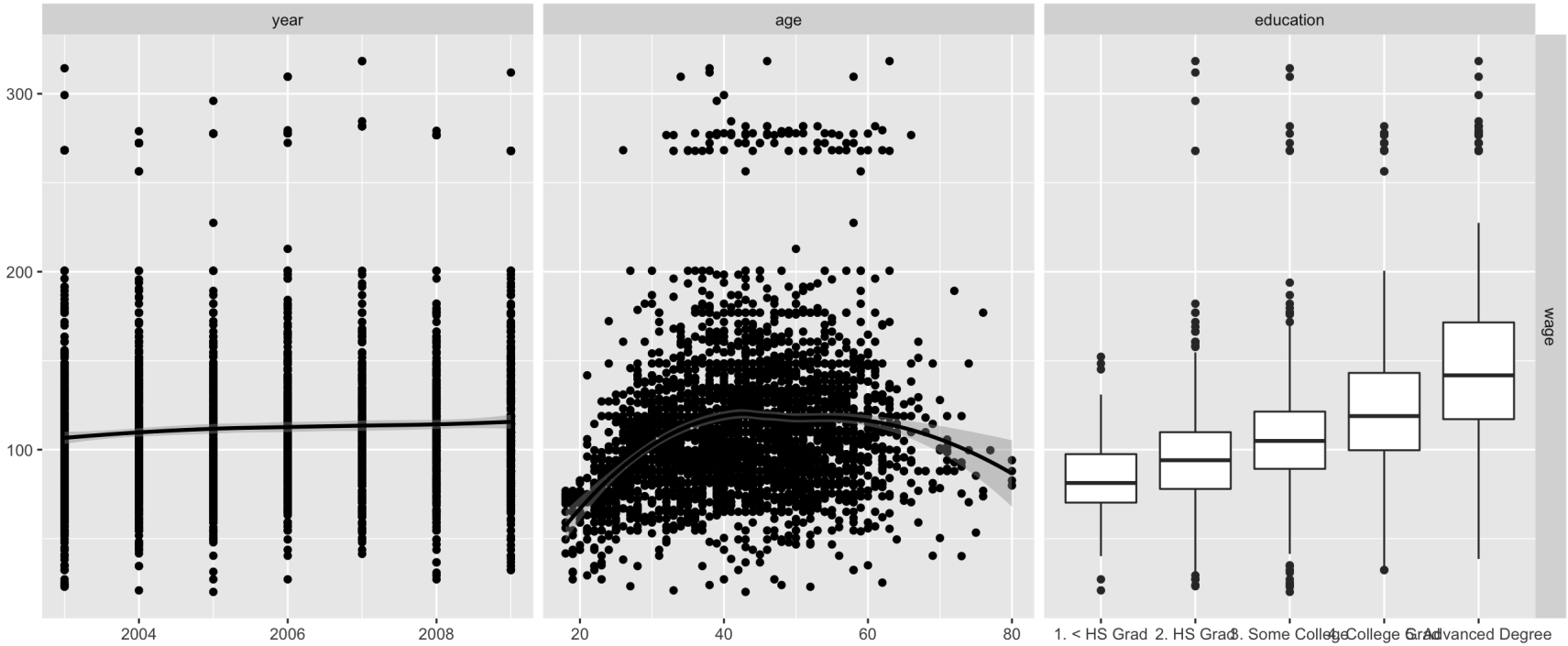
# Example - Wages

Wage and other data for a group of 3000 male workers in the Mid-Atlantic region. Goal: Predict wage based on worker characteristics.

```
## Rows: 3,000
## Columns: 11
## $ year       <int> 2006, 2004, 2003, 2003, 2005, 2008, 2009, 2008, 2006, 2004
## $ age        <int> 18, 24, 45, 43, 50, 54, 44, 30, 41, 52, 45, 34, 35, 39, 54
## $ maritl     <fct> 1. Never Married, 1. Never Married, 2. Married, 2. Married
## $ race       <fct> 1. White, 1. White, 1. White, 3. Asian, 1. White, 1. White
## $ education  <fct> 1. < HS Grad, 4. College Grad, 3. Some College, 4. College
## $ region     <fct> 2. Middle Atlantic, 2. Middle Atlantic, 2. Middle Atlantic
## $ jobclass   <fct> 1. Industrial, 2. Information, 1. Industrial, 2. Informati
## $ health     <fct> 1. <=Good, 2. >=Very Good, 1. <=Good, 2. >=Very Good, 1. <
## $ health_ins <fct> 2. No, 2. No, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Y
## $ logwage    <dbl> 4.318063, 4.255273, 4.875061, 5.041393, 4.318063, 4.845098
## $ wage       <dbl> 75.04315, 70.47602, 130.98218, 154.68529, 75.04315, 127.11
```

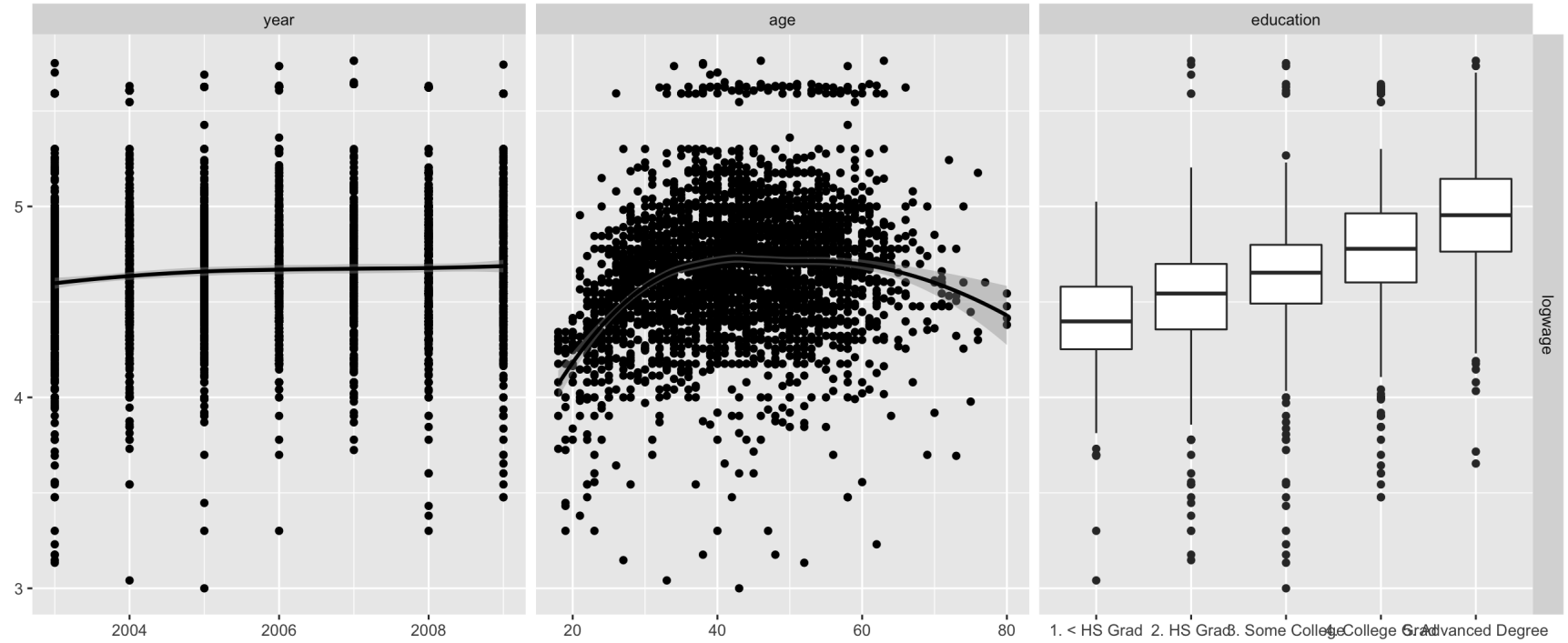
# Take a look

What do the following pairwise comparisons of the variables `year`, `age` and, `education` against `wage` show us?



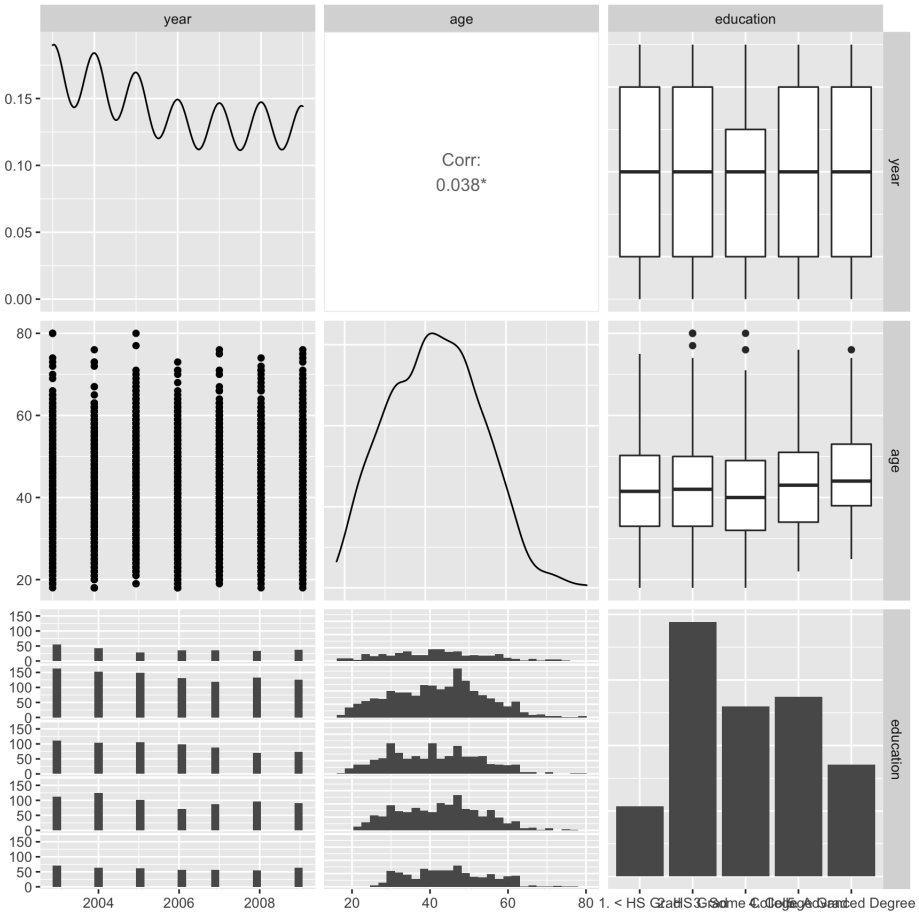
# Take a look

If we examine `logwage` instead of `wage` as response variable - what changes?



# Take a look

- Examine the predictors. *Ideally values are spread out without any associations.*
- There is no evidence of any association between the predictors here.
- One category of "education" has much fewer observations than the others - this is not ideal.



# Model for wage data

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

where $Y =$ `logwage`, $X_1 =$ `year`, $X_2 =$ `age`, and $X_3 =$ `education`.

# Fitting the model in R

```
library(broom)
library(parsnip)
library(kableExtra)
lm_mod <-
  linear_reg() %>%
  set_engine("lm")

fit <- lm_mod %>%
  fit(logwage~year+age+education, data=Wage)
```

# Parameter estimates

```
tidy(fit) %>%
  kable(digits = 3) %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | -17.450 | 5.469 | -3.191 | 0.001 |
| year | 0.011 | 0.003 | 3.952 | 0.000 |
| age | 0.006 | 0.000 | 11.447 | 0.000 |
| education2. HS Grad | 0.120 | 0.021 | 5.762 | 0.000 |
| education3. Some College | 0.244 | 0.022 | 11.115 | 0.000 |
| education4. College Grad | 0.368 | 0.022 | 16.894 | 0.000 |
| education5. Advanced Degree | 0.541 | 0.024 | 22.909 | 0.000 |

# Model fit

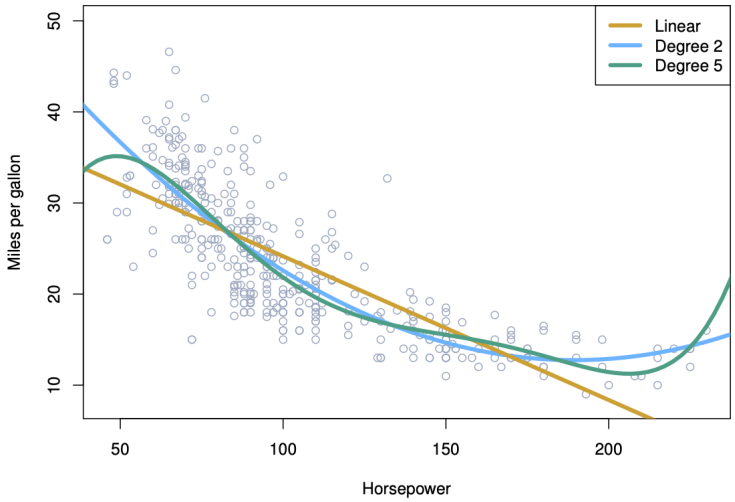
```
glance(fit)  %>%
  kable(digits = 2) %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

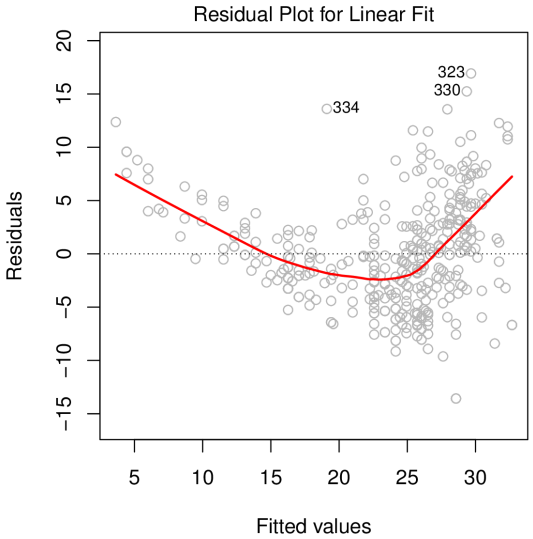
| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residual | nobs |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.26 | 0.26 | 0.3 | 178.09 | 0 | 6 | -663.9 | 1343.8 | 1391.85 | 273.44 | 2993 | 3000 |

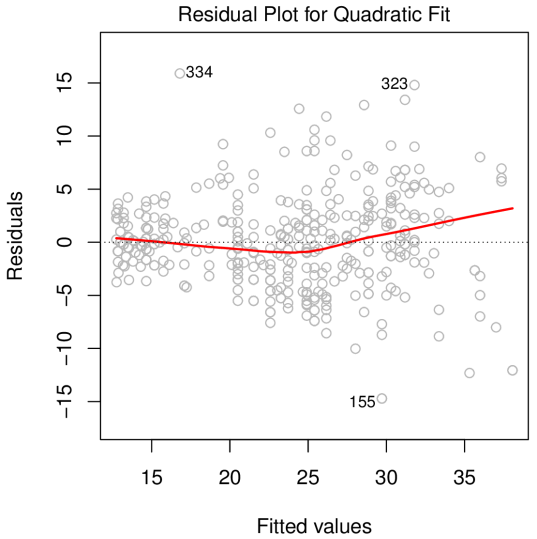
# Polynomial regression

A simple non-linear model can be achieved by adding polynomial terms, like

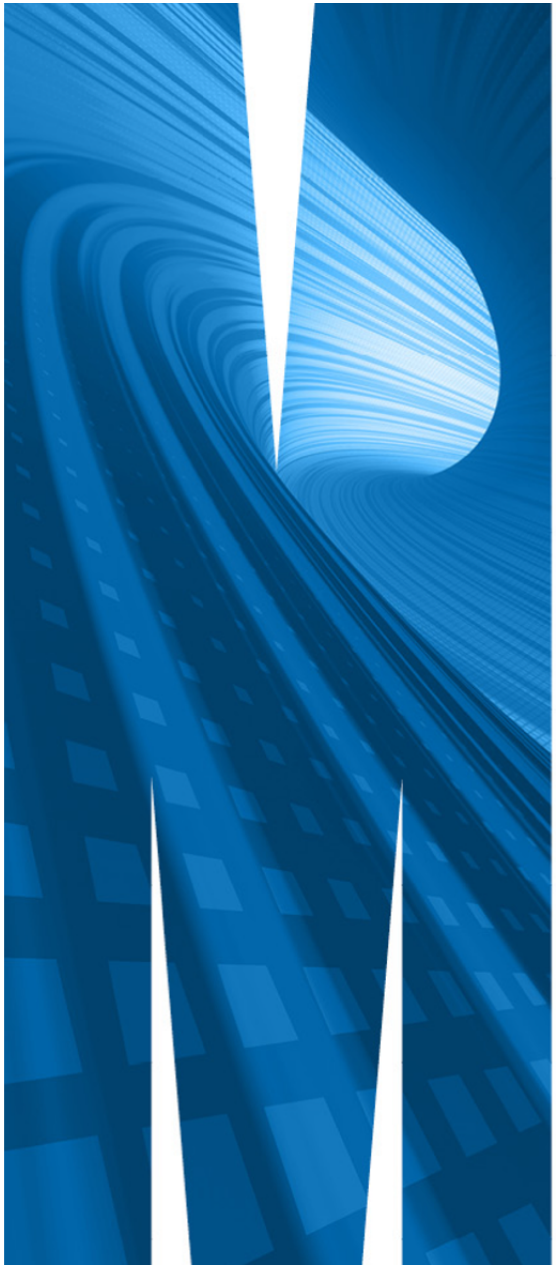
$$\text{mpg} = \beta_0 + \beta_1 \text{horsepower} + \beta_2 \text{horsepower}^2 + \varepsilon$$




(Chapter3/3.8.pdf,3.9.pdf)





Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 2a