UNIVERSITAT DE BARCELONA

Linear Regression

Research Methods in Cyberspace, Behavior and e-Therapy

David Leiva Ureña

dleivaur@ub.edu

Quantitative Psychology Section. Faculty of Psychology.

Simple Linear Regression

Introduction

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Some basic terminology and notions related to statistical modeling:
  - A **model** is a formal representation of a real phenomenon. In this course we have used statistical models to describe reality.
  - That variable to be predicted is the **response** variable, also called dependent, endogen or outcome variable (Y).
  - Those set of variables used to make predictions about the response are the **predictors**. These are also called regressors, independent or exogen variables ($X_1, X_2, \ldots, X_p$).
  - The statistical procedure for modeling the relationship between a response variable and a set of predictors is known as **regression** analysis.
  - It is important to note that we usually deal with observational data, thus despite we establish a functional relationship it is done based on the association pattern. To establish causality we need to rely on experimental data.
  - Mathematical notation:

$$\begin{aligned} Y &= E(Y) + \epsilon \\ E(Y) &= f(X) \\ E(Y) &= X_1 + X_2 + \cdots + X_p \end{aligned}$$

Simple Linear Regression

Introduction

Simple Linear Regression

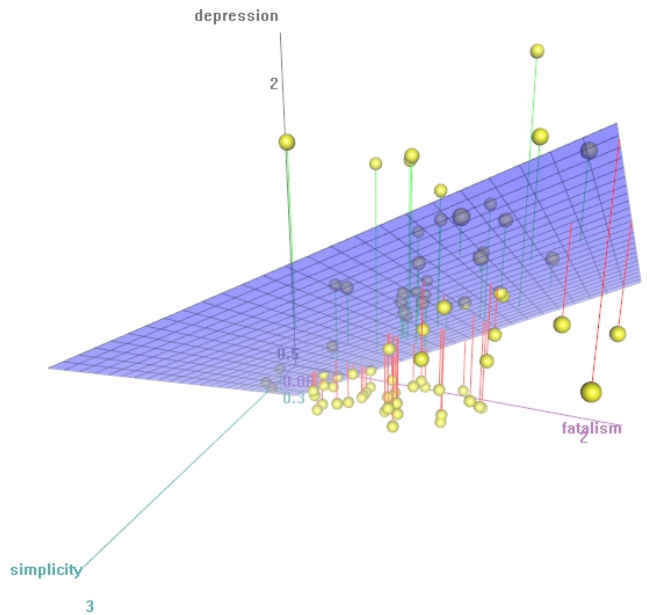
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Main steps in regression models:
  - Model specification
  - Data collection
  - Parameters estimation
  - Model validation
  - Utility of the model
  - Prediction and estimation



Simple Linear Regression

Specification of the model

Simple Linear Regression

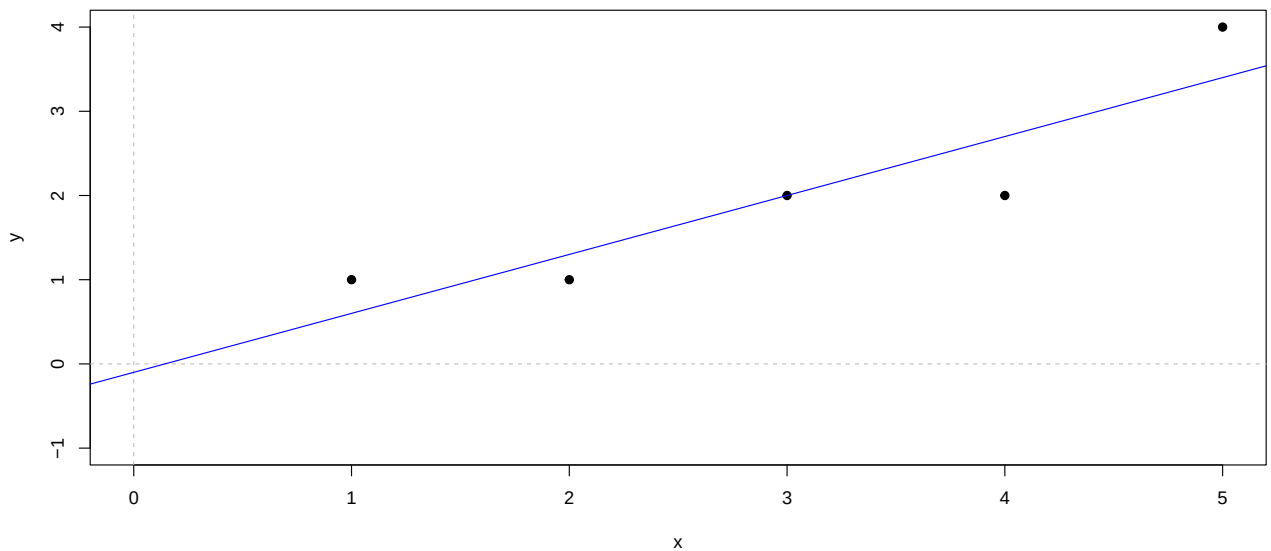
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The simplest model establishes that the response is a function of one predictor: $y = f(x) + \epsilon$.
- Given that the relationship is assumed to be linear, $f(x)$ represents a line.
- Thus, $y = \beta_0 + \beta_1 x + \epsilon$.



Simple Linear Regression

Specification of the model

Simple Linear Regression

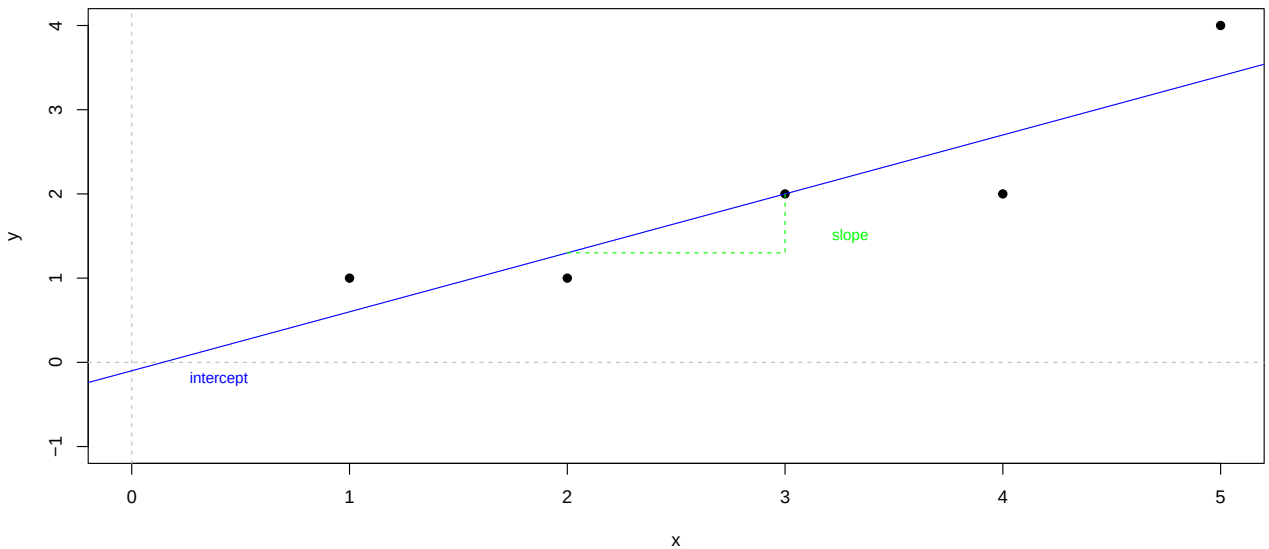
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The linear model is formed by a deterministic part $\beta_0 + \beta_1 x$, that is to say, the expected value of y ($E(y)$), and a random one ϵ, or the error term of the model.
- β_0 is the y-intercept of the line, also known as the constant of the model.
- β_1 is the slope of the line, and that can be positive or negative.



Simple Linear Regression

Specification of the model

Simple Linear Regression

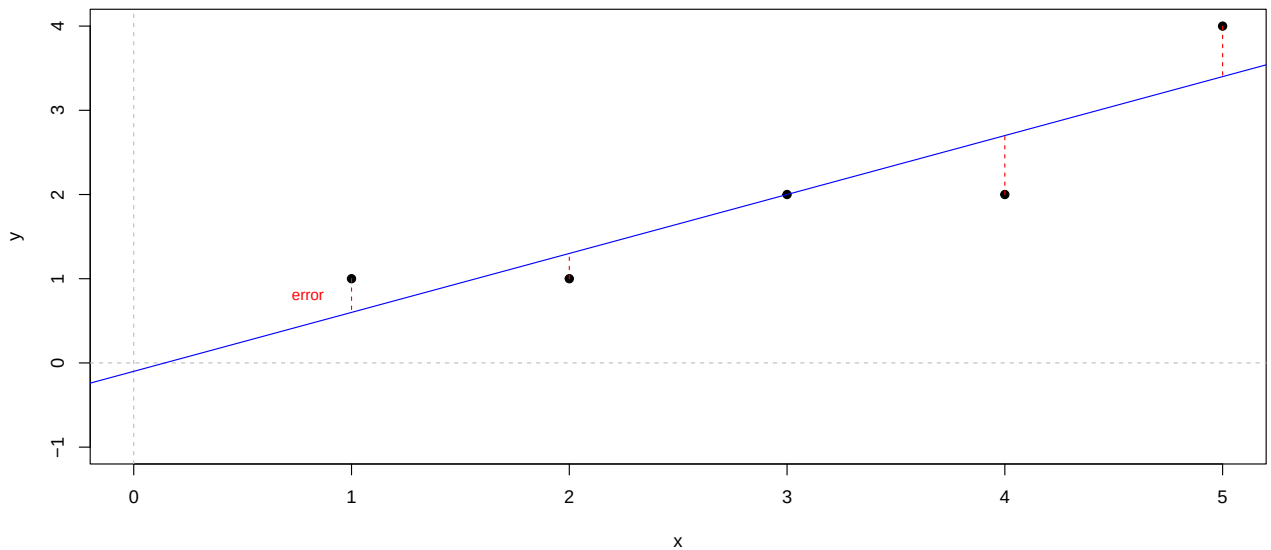
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Let define the error for the ith observation as the difference between this observation and its predicted value according the fitted model, $y_i - \hat{y}_i$.
- The predicted value of y is: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Thus $y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$.



Simple Linear Regression

Parameters Estimation

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Modern statistical software have several routines to estimate model's parameters, the most well-known are those based on **Ordinary Least Squares**.
- Fitting linear model by OLS implies to find estimates of β_0 and β_1 that: $\min\left(\sum_{i=1}^{n} e_i^2\right) = min\left(\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2\right)$.
- The least squares line satisfies:
  - SE=$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$.
  - SSE = $\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$ is minimum.
- After some calculus we obtain:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- The ordinary least squares estimates are the best linear unbiased estimators (BLUE) under the assumptions of the linear model.

Simple Linear Regression

Model's Assumptions

Simple Linear Regression

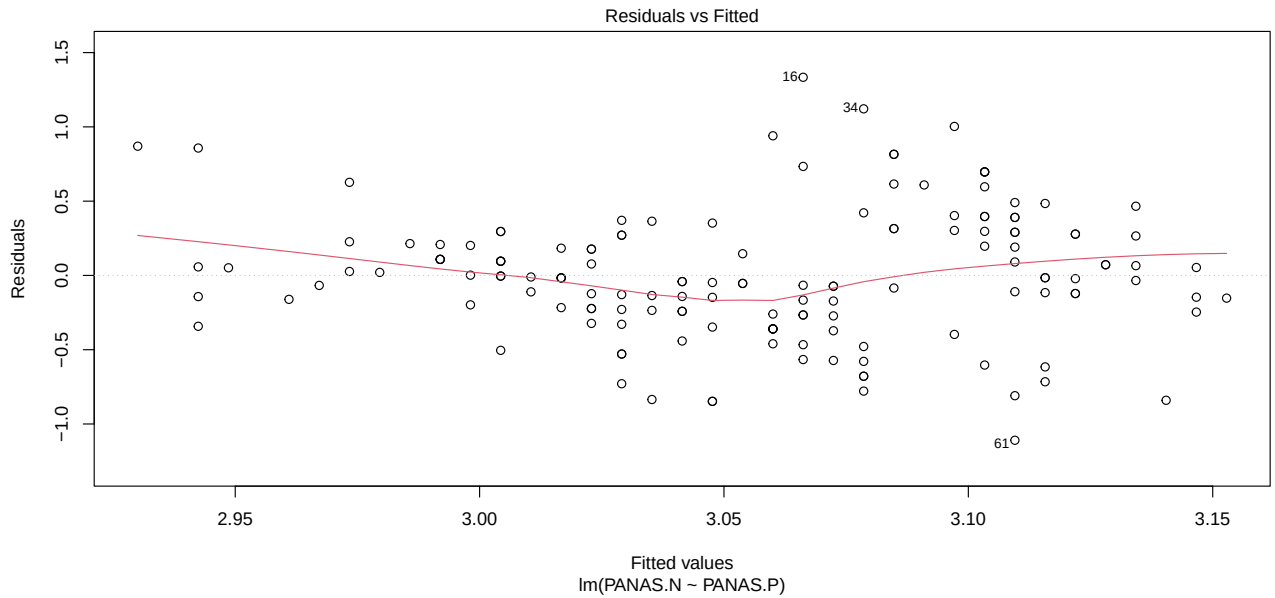
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- **Linearity assumption**: The mean of the probability distribution of ϵ is 0 $\rightarrow E(\epsilon) = 0$.
- It is logical since we restricted the possible solutions when fitting the model to those that met SE=0.
- It also implies that $E(y) = \beta_0 + \beta_1 x$.



Simple Linear Regression

Model's Assumptions

Simple Linear Regression

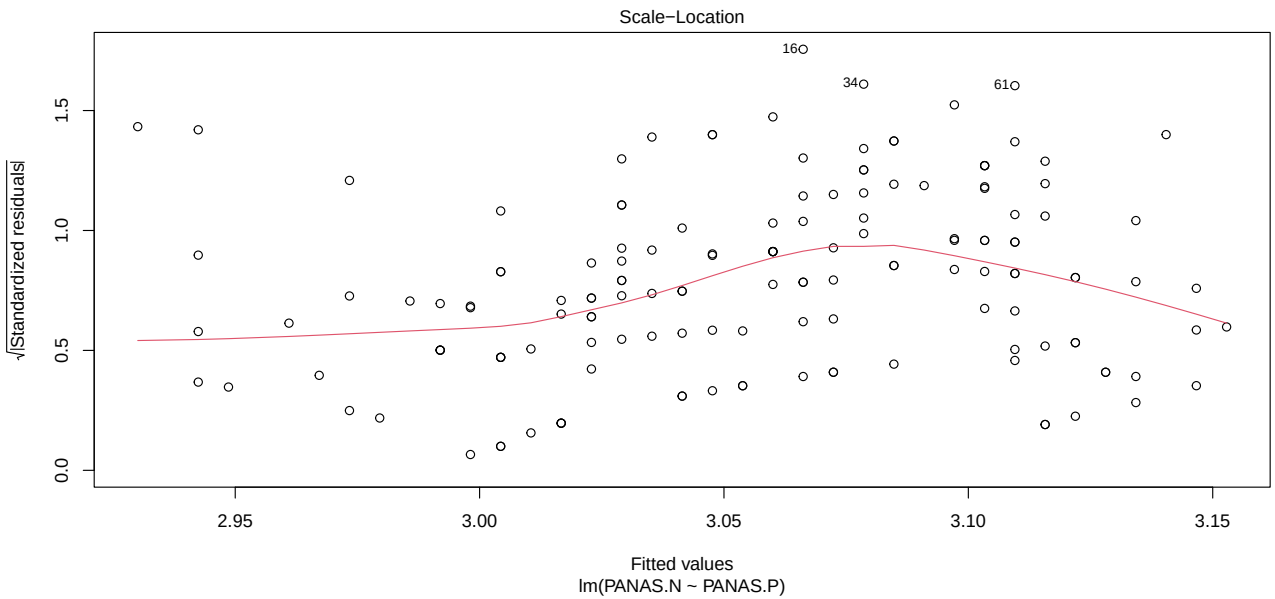
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- **Homoscedasticity assumption**: The variance of the probability distribution of ϵ is constant for all values of the predictor (x). That implies that $\sigma^2_\epsilon = \sigma^2$, where σ^2 is a constant.
- In other words, the variance of the error does not depend on the x values.



Simple Linear Regression

Model's Assumptions

Simple Linear Regression

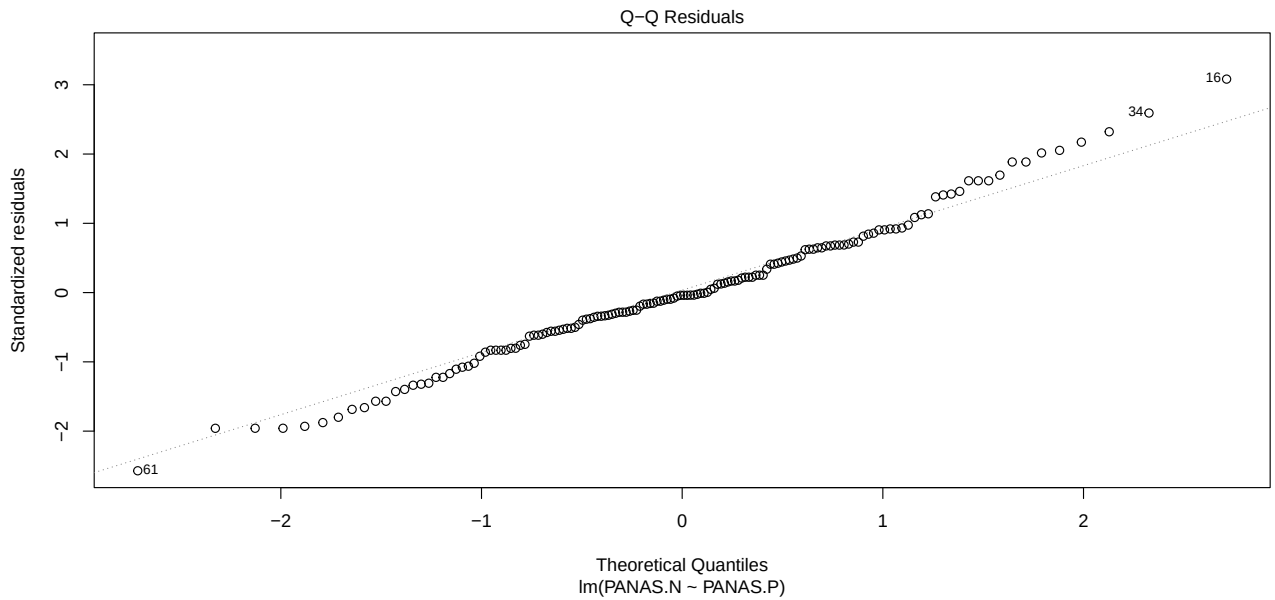
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- **Normality assumption**: Taking into account the previous assumptions, the probability distribution of the errors corresponds to a normal distribution with $\mu = 0$ and $\sigma^2 = \sigma^2_\epsilon$. That is, $\epsilon \sim N(0, \sigma^2_\epsilon)$.



Simple Linear Regression

Model's Assumptions

Simple Linear Regression

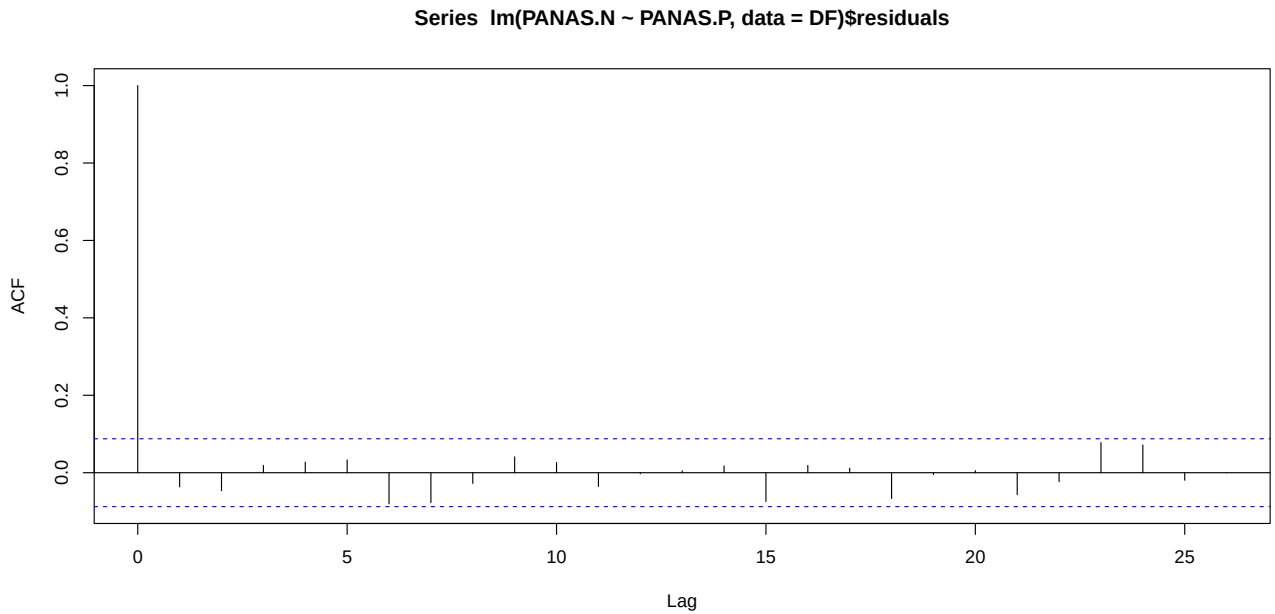
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- **Independence assumption** The errors associated with two different observations are independent.
- Depending on the design of the research (or the experiment) we can assume independent errors.
- If we cannot assume stochastic independence amongst errors, OLS estimates are biased.



Simple Linear Regression

Inference with the Regression Model

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The variability of the random error σ^2 is related to the utility of the fitted model. The greater the variability of the random error, the poorer the fitted model in terms of the estimation of the parameters and the prediction.
- σ^2 is unknown and the best estimator is s^2 which can be obtained as follows:

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

- Some statistical packages provides with the so-called Root of Mean Square Error: RMSE = $s = \sqrt{s^2}$.
- We expect that most of the observations (about 95%) lie within $\hat{y} \pm 2 \times RMSE$.[a]

[a] Due to normality of random error assumption.

Simple Linear Regression

Inference with the Regression Model

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- We are interested in testing the utility of the predictor in the model.
- The statistical hypothesis in this case is as follows:

$$H_0 : \beta_1 = 0$$

- If the assumptions of the model are met then the sampling distribution of $\hat{\beta}_1$ is $N(\mu = \beta_1, \sigma_{\hat{\beta}_1} = \frac{\sigma_\epsilon}{\sqrt{n}s_x})$.
- The test statistic in this case is Student's t:

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{RMSE/\sqrt{n}s_x} \sim t_{n-2}$$

- A confidence interval for the slope, β_1, can be obtained as follows:

$$\hat{\beta}_1 \pm t_{\nu,\alpha/2} \times s_{\hat{\beta}_1}$$

Simple Linear Regression

Inference with the Regression Model

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Another possibility is making statistical decisions regarding the intercept. Therefore, we decide whether the regression line goes through the origin.

$$H_0 : \beta_0 = 0$$

- If the assumptions of the model are met then the sampling distribution of $\hat{\beta}_0$ is $N\left(\mu = \beta_0, \sigma_{\hat{\beta}_0} = \frac{\sigma_\epsilon}{\sqrt{n}}\sqrt{\left[1 + \frac{\bar{x}^2}{s_x^2}\right]}\right)$.
- The test statistic in this case is Student's t:

$$t = \frac{\hat{\beta}_0}{s_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{\frac{RMSE}{\sqrt{n}}\sqrt{\left[1 + \bar{x}^2/s_x^2\right]}} \sim t_{n-2}$$

- A confidence interval for the intercept, β_0, can be obtained as follows:

$$\hat{\beta}_0 \pm t_{\nu,\alpha/2} \times s_{\hat{\beta}_0}$$

Simple Linear Regression

Predictive Capacity of the Model

Simple Linear Regression

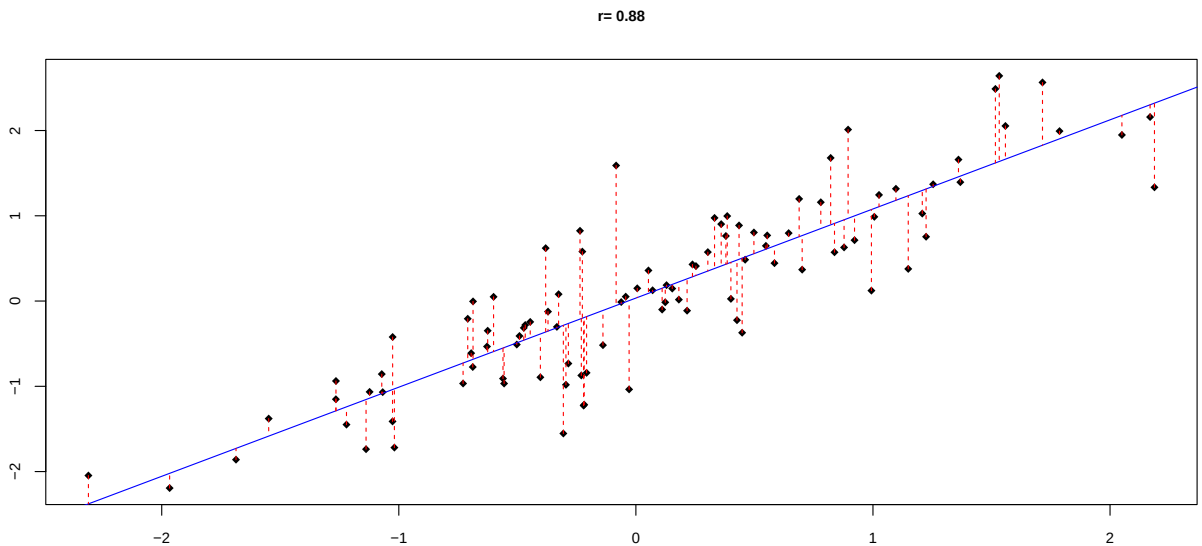
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The predictive capacity of a regression model is interpreted as that part of the variability in the response that this model can account for.
- Let's plot the errors of the model including the predictor:



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$SSE = 27.12$$

Simple Linear Regression

Predictive Capacity of the Model

Simple Linear Regression

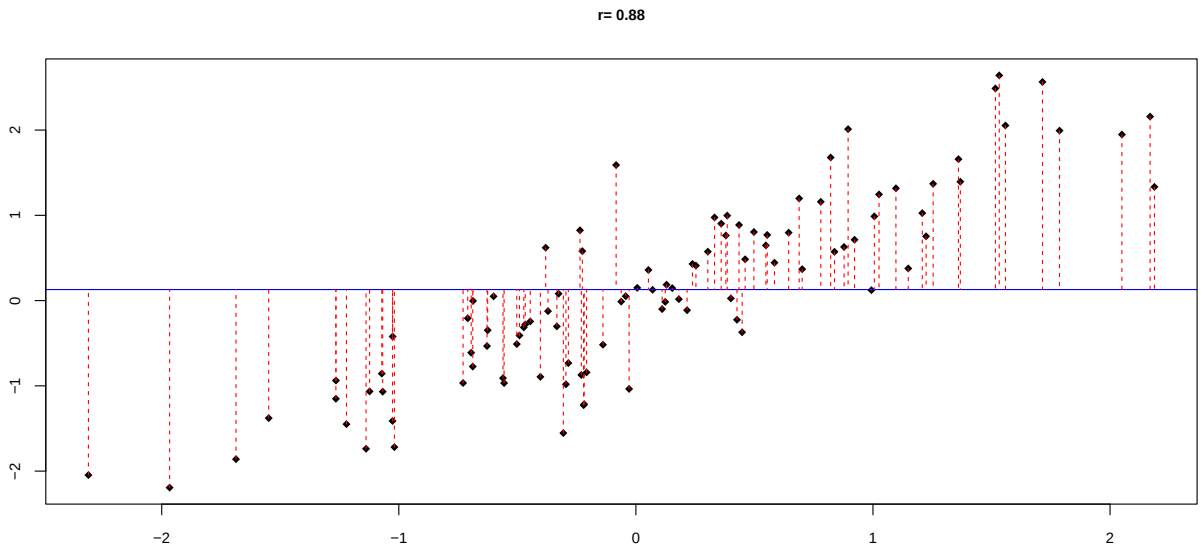
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The predictive capacity of a regression model is interpreted as that part of the variability in the response that this model can account for.
- Now let's plot the errors for a null model, that is, a model without predictors:



$$\hat{y}_i = \hat{\beta}_0$$

$$SSY = 117.24$$

Simple Linear Regression

Predictive Capacity of the Model

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Thus, by means of the coefficient of determination we quantify the capacity of the model to predict the response.

$$R^2 = 1 - \frac{SSE}{SSY}$$

- It ranges from 0 to 1 though is usually expressed as a percentage.
- R^2 can be interpreted as the percentage of the response variable that can be explained by the model.
- In the case of the simple regression model $R^2 = r_{xy}^2$

Simple Linear Regression

Prediction: Mean Response

Simple Linear Regression

Models diagnostics

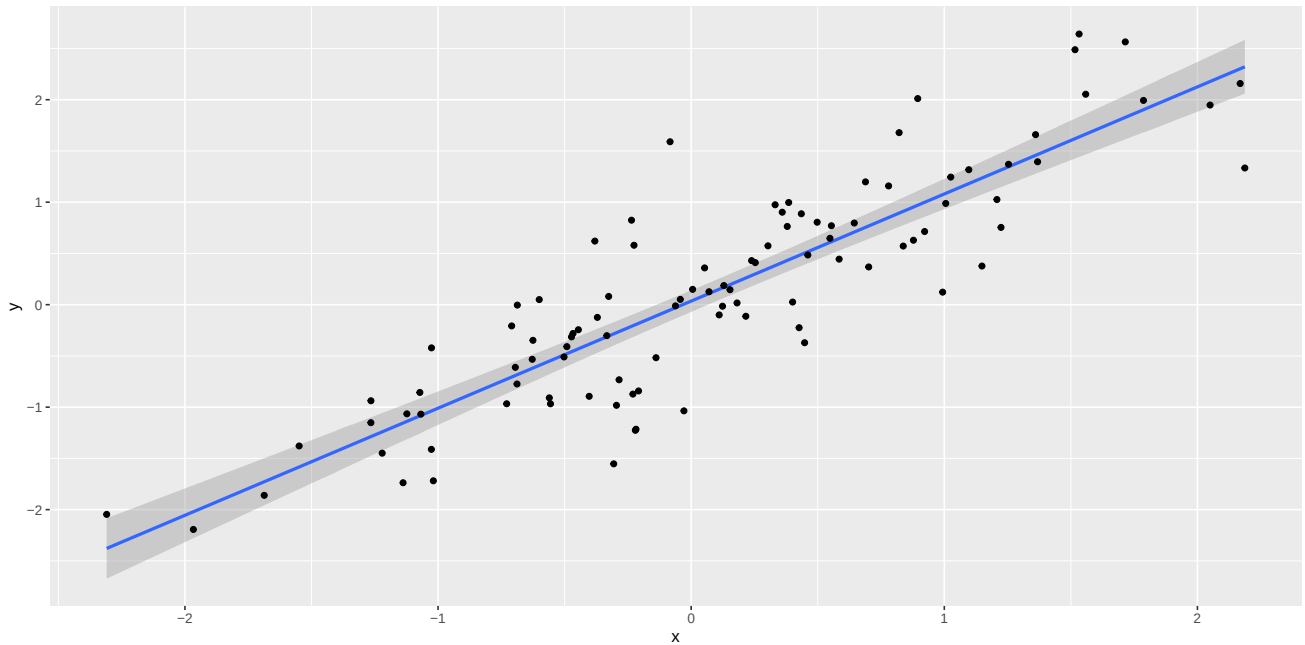
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Estimating the mean value of y given a specific value of x.
- Confidence interval for the mean value of y:

$$\hat{y} \pm t_{\nu,\alpha/2} \times s \times \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_x^2}}$$



Simple Linear Regression

Prediction: Individual Response

Simple Linear Regression

Models diagnostics

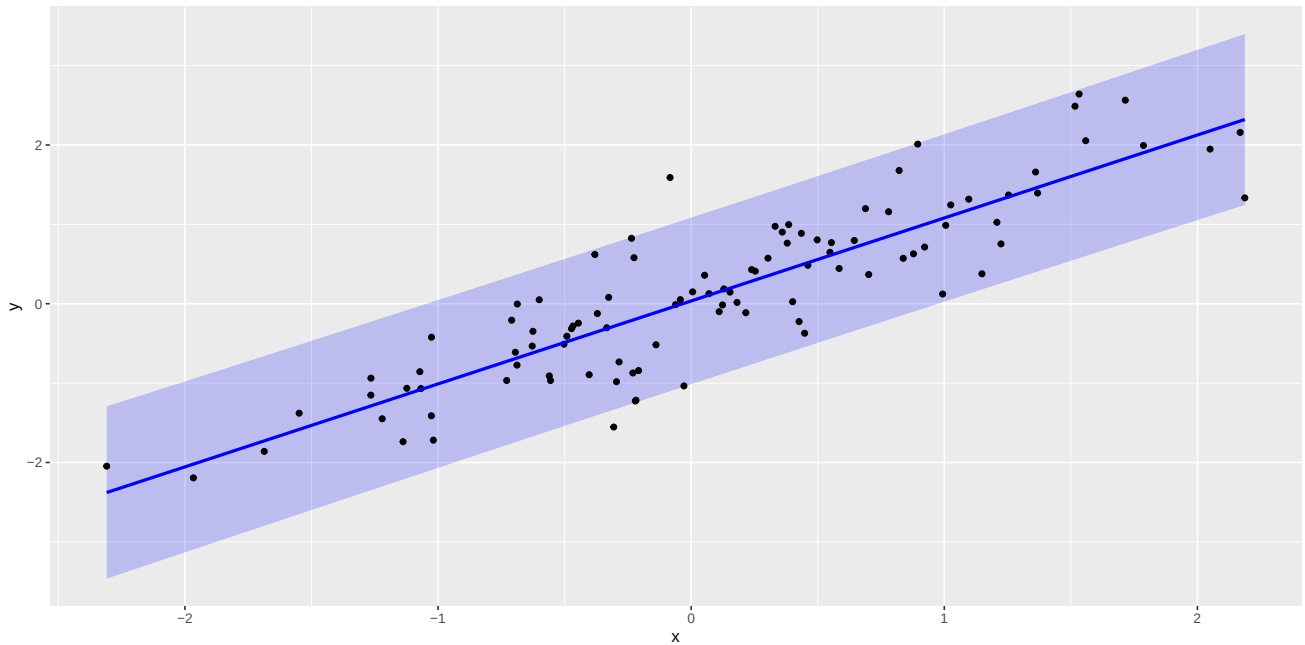
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Predicting a new observation given a specific value of x.
- Prediction interval for an individual value of y:

$$\hat{y} \pm t_{\nu,\alpha/2} \times s \times \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_x^2}}$$



Simple Linear Regression

Example

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Social prestige as a function of income.

	Dependent variable:
	prestige
income	0.003***
	(0.0003)
Constant	27.141***
	(2.268)
Observations	102
R^2	0.511
Adjusted R^2	0.506
Residual Std. Error	12.090 (df = 100)
F Statistic	104.537*** (df = 1; 100)
Note:	*p<0.1; **p<0.05; ***p<0.01

Simple Linear Regression

Example

Simple Linear Regression

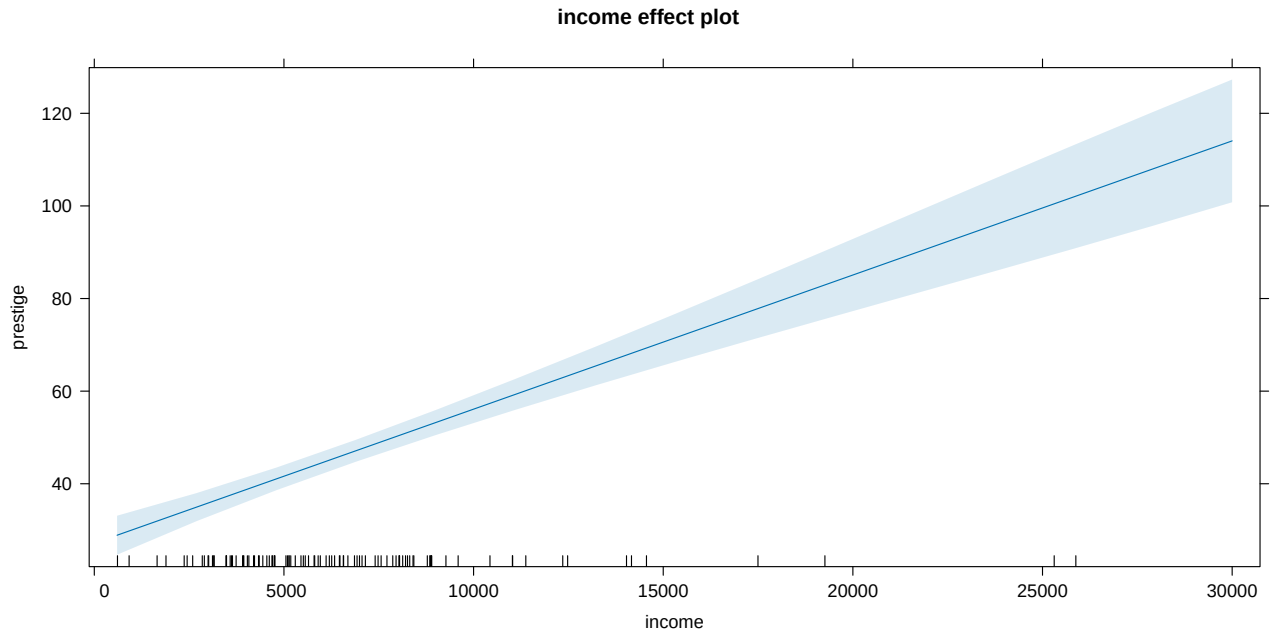
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Social prestige as a function of income.



Linear Models Diagnostics

Assumptions: Linearity

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- Linear relationship is assumed between predictors and response variable.
- If the model is adequately specified we can assume that $E(\epsilon) = 0$. This implies that: $E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.
- Warning: A misspecified model can yield significant results when assessing its utility (e. g., by means of an F-test).
- Residuals plots and RESET test are suitable for detecting model's lack of fit.
- RESET test allows to decide whether the linear model is adequately specified.
- The null hypothesis states that the linear model is appropriately specified. The alternative hypothesis implies that adding a high order term in the model will improve its fit.
- Recall the importance of the balance between model's fit and model's parsimony.

Linear Models Diagnostics

Assumptions: Linearity

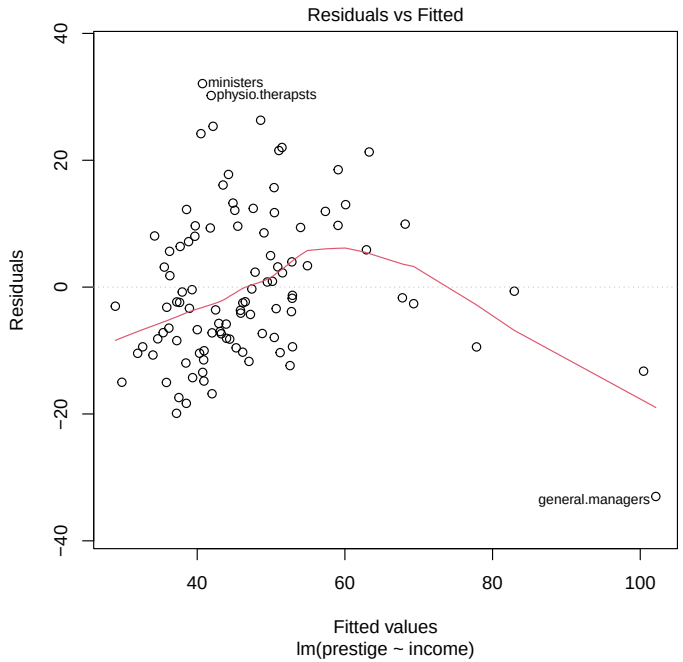
Simple Linear Regression

Models diagnostics

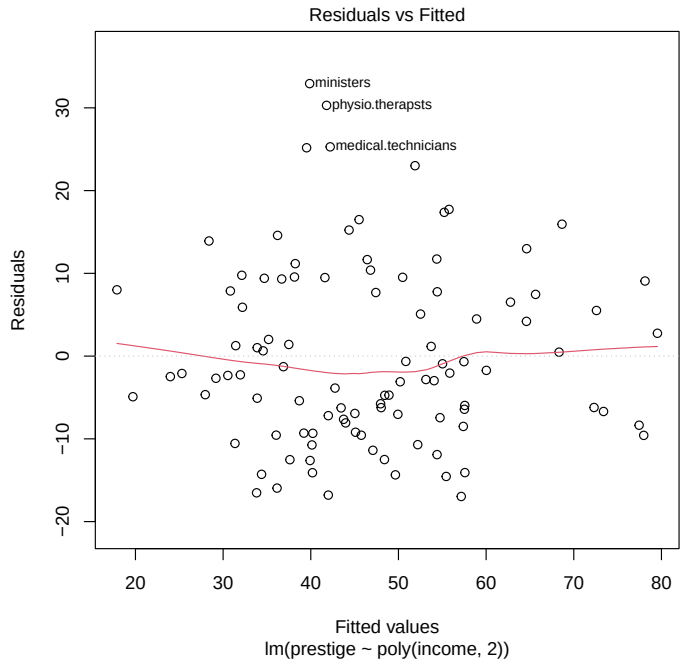
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



Linear model



Polynomial model

Linear Models Diagnostics

Assumptions: Homoscedasticity

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- This assumption implies that the variability of residuals remains constant through all levels of the predictors.
- When equal variances can be assumed we called it homoscedastic. On the contrary, heteroscedastic variances occur when these are unequal.
- Departures from homoscedasticity can be detected by means of plots and statistical tests (Breusch-Pagan test).
- If heteroscedasticity is found, there are several options for correcting it (for instance, transformations or using weighted least squares estimates).
- Breusch-Pagan's test allows to decide whether the variances can be assumed to be unequal.
- The null hypothesis states that the variances are equal. Thus, its rejection implies assuming heteroskedasticity.

Linear Models Diagnostics

Assumptions: Homoscedasticity

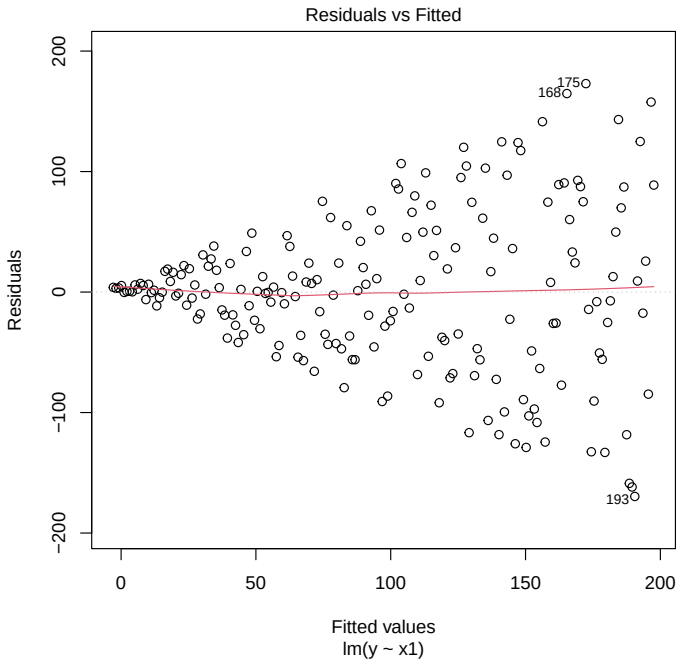
Simple Linear Regression

Models diagnostics

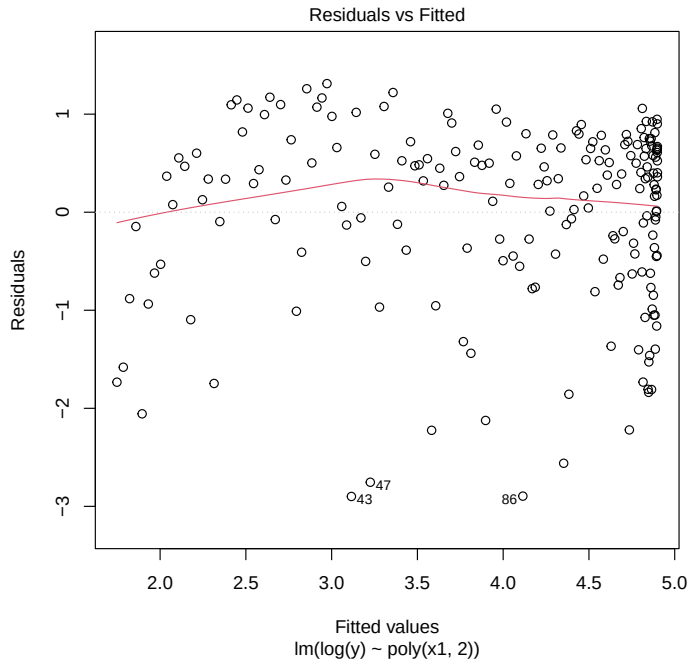
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



$y \sim x_1$



$\log(y) \sim x_1 + x_1^2$

Linear Models Diagnostics

Assumptions: Homoscedasticity

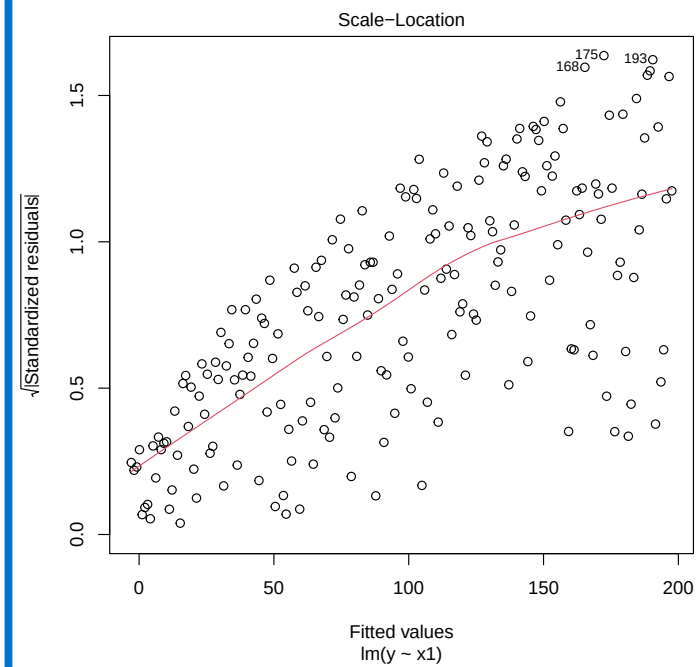
Simple Linear Regression

Models diagnostics

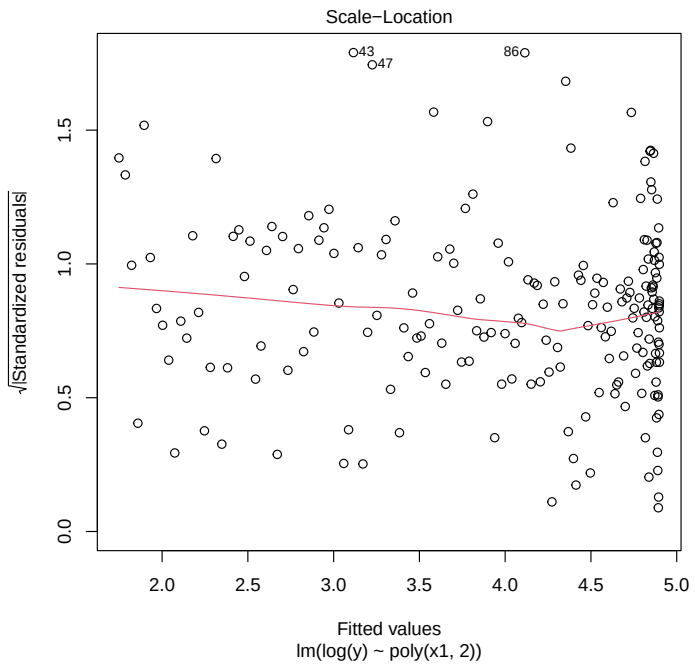
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



$y \sim x_1$



$\log(y) \sim x_1 + x_1^2$

Linear Models Diagnostics

Assumptions: Normality

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The errors are assumed to be normal in the population Normal$(0, \sigma^2)$.
- If the normality assumption is violated the estimates of the linear model will be biased. Nevertheless, small deviations from the normal distributions are not problematic for linear models.
- In order to check this asssumption we can use statistical tests as Kolmogorov-Smirnov or Shapiro-Wilk tests.
- Normal quantile plots are useful to this aim as well.
- Shapiro-Wilk tests is more powerful than other normality tests (e. g., Kolmogorov-Smirnov test). It is specially suitable for small sample sizes ($n < 30$).
- Nevertheless, given the rationale of the statistical test (and the well-known relationship between the p-value and sample size), most statisticians do prefer using quantile plots to assess deviations from normality.

Linear Models Diagnostics

Assumptions: Normality

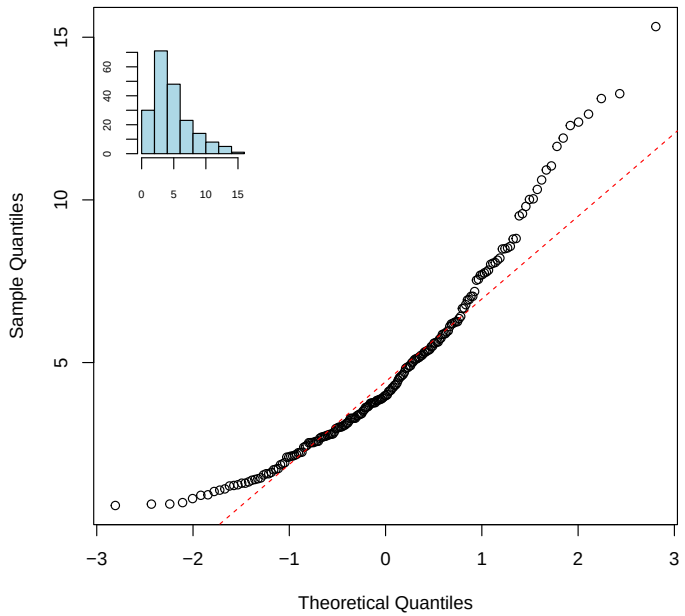
Simple Linear Regression

Models diagnostics

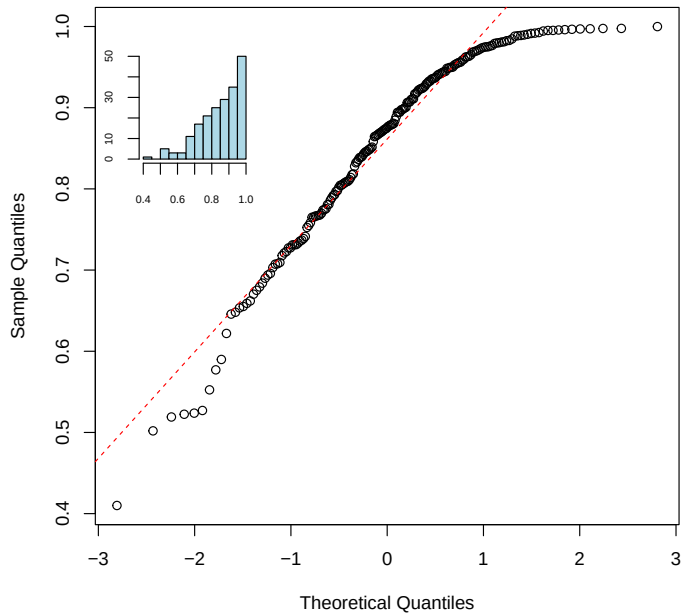
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



Right-skewed: Chi-squared distribution



Left-skewed: Beta distribution

Linear Models Diagnostics

Assumptions: Normality

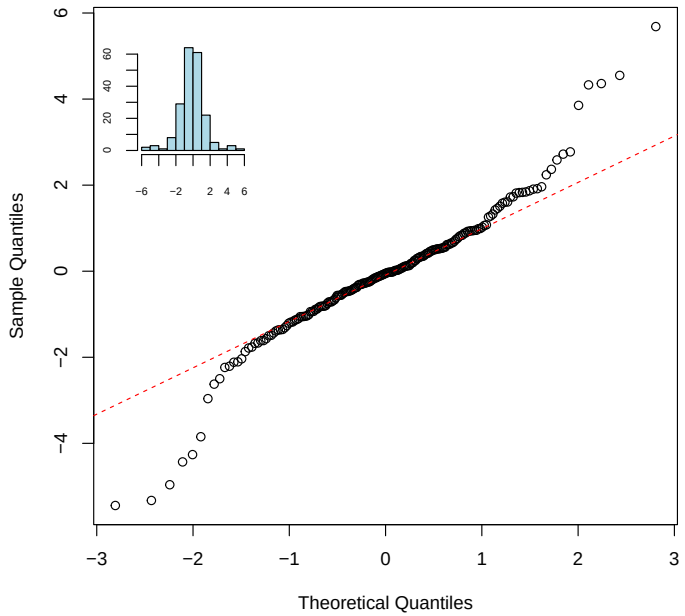
Simple Linear Regression

Models diagnostics

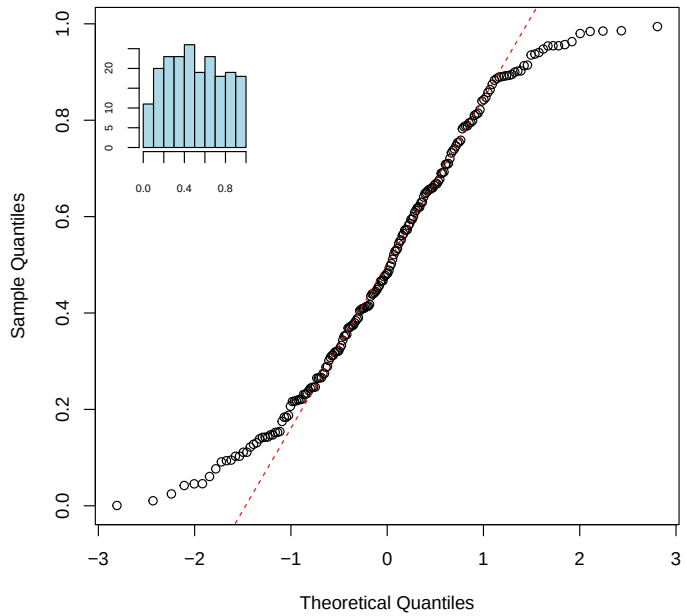
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



Light-tailed distribution: t



Heavy-tailed distributions: Uniform

Linear Models Diagnostics

Assumptions: Independence

Simple Linear Regression

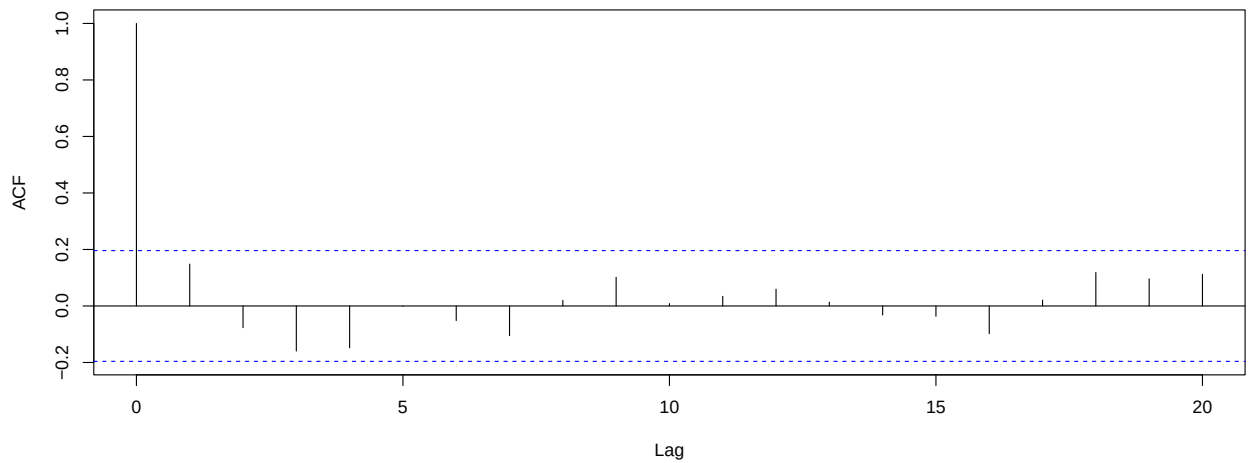
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- This assumption implies that the covariance (so the correlation as well) between the residuals of any pair of observations is 0.
- If dealing with studies in which independent observations are obtained regarding a set of random variables, the design of the study could be enough to assume the independenee between errors in the population.
- Again, some statistical tests and plots might be used if we were interested in checking the independence assumption.
- Durbin-Watson test of autocorrelation: $H_0 : \rho = 0$.
- Plot of the autocorrelation function.



Linear Models Diagnostics

Outliers and Influential Observations

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- An outlier is an extreme observation. In the case of linear models, it corresponds to a residual extremely large.
- Different types of residuals:
  - Standardized residuals: $z_i = \frac{e_i}{s} = \frac{y_i - \hat{y}_i}{s}$. z_i greater than $3s$ is considered an outlier.
  - Studentized residuals: $z_i^* = \frac{e_i}{s\sqrt{1-h_i}} \sim t_{n-p-1}$.
  - Deleted residuals: $d_i = y_i - \hat{y}_{(i)}$.
  - Studentized deleted residuals: $d_i^* = \frac{d_i}{SE(d_i)} \sim t_\nu$.
- Causes of outliers: measurement or coding errors, non-homogeneous samples...
- Note that the previous types of outliers are all referred to the response. But we can also observe outliers in the predictors, this is related to the so-called **leverage**.
- The leverage (also called the influence h_i) is the weight of a given observation over the fitted value. In this case, the observation corresponds to the observed value for a predictor or a set of predictors.

Linear Models Diagnostics

Outliers and Influential Observarions

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The larger the leverage the greater the influence over the predicted value.
- In general, $h_i > \frac{2(k+1)}{n}$ is considered influential.
- In order to study the impact of the observations over the model we use other influence indices that combines both sources of variability (response and predictors).
- A commonly used measure is the *Cook's distance*, which is a measure of real influence of the observation over the estimated coefficients of the model.

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1)\text{MSE}} \frac{h_i}{(1-h_i)^2}$$

- D_i can be compared to $F_{k+1;n-(k+1)}$. But it's rather conservative.
- Other rule of thumb is considering influential observations those $D_i > 4/n$.
- Other real influence indicators are: Dffits and Dfbetas. This indices can be computed with R syntax.

Linear Models Diagnostics

Outliers and Influential Observarions

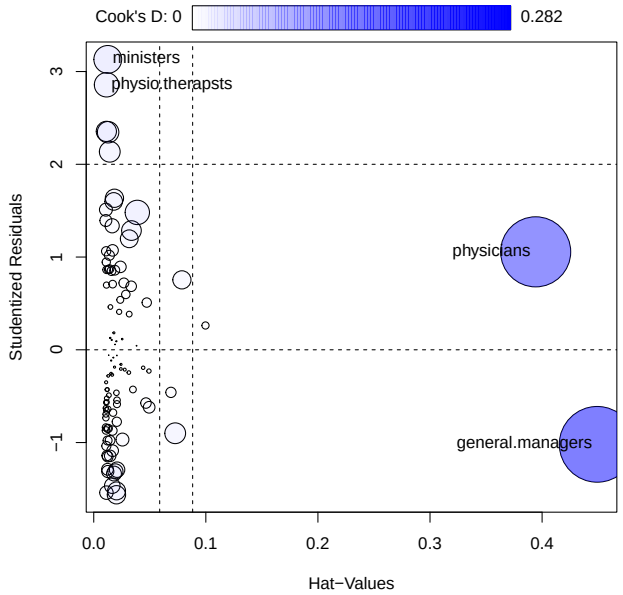
Simple Linear Regression

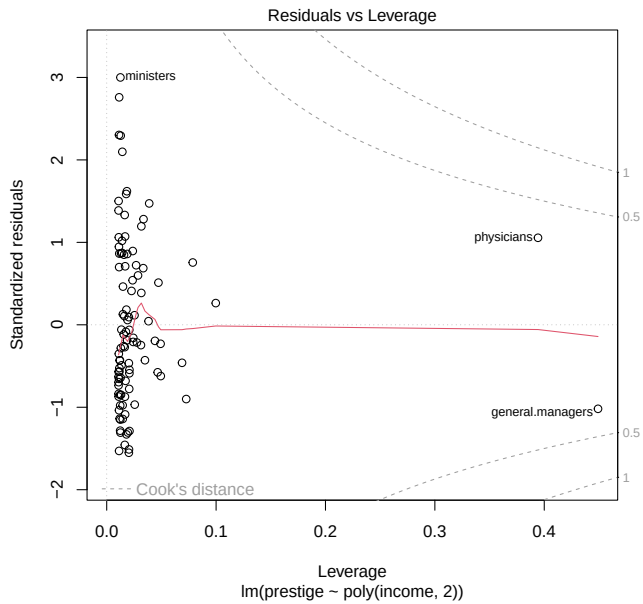
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models





Multiple Linear Regression Models

Introduction

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Models that include more than one predictor are called multiple regression models.
- Thus, the dependent variable y is a linear function of several independent variables $x_1, x_2, \ldots, x_k$.
- The deterministic part of the model is complemented by the random component ϵ.
- Mathematical notation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

- Same steps as those presented for the simple model: Specification, estimation (and inference),validation, and prediction.
- The linear model is formed by a deterministic part $\beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_k x_k$, that is to say, the expected value of y ($E(y)$), and a random one ϵ, or the error term of the model.
- β_0 is the y-intercept of the model.
- β_i is the partial contribution of each of the k predictors.

Multiple Linear Regression Models

Introduction

Simple Linear Regression

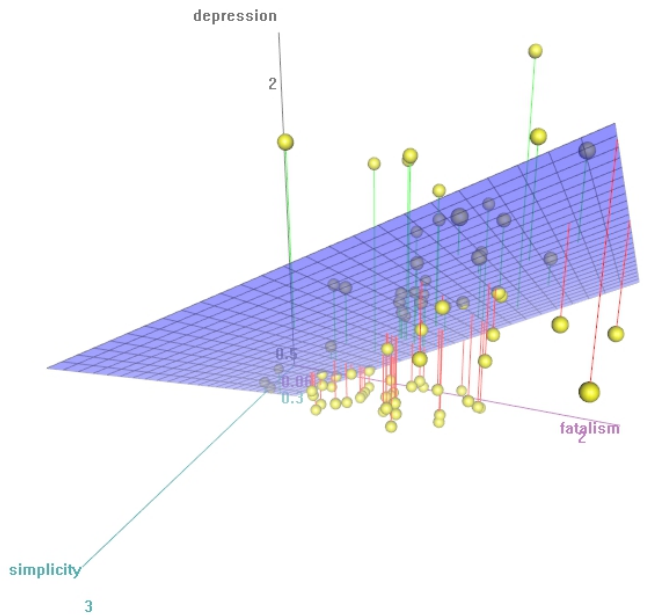
Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The model is $y = f(x_1 + x_2 + \cdots + x_k) + \epsilon$.
- The relationship is assumed to be linear, $f(x_1 + x_2 + \cdots + x_k)$, we fit therefore a plane or an hyperplane, depending on the number of regressors.



Multiple Linear Regression Models

Estimation

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Let define the error for the ith observation as the difference between this observation and its predicted value in the fitted model, $y_i - \hat{y}_i$.
- The predicted value of y is: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}$.
- Thus $y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}\right)$.
- Fitting linear model implies to find estimates of β_is that accomplish: $\min\left(\sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}\right)\right]^2\right)$.
- We need matrix algebra to solve this problem:

$$y = \mathbf{X}\beta$$
$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$$

- The ordinary least squares estimates are the best linear unbiased estimators (BLUE) under the assumptions of the linear model.
- Model's assumptions: Linearity, normality, homoscedasticity, and independence.

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- We are interested in testing the utility of some of the predictors in the model. Thus, β_i for $i = 1, 2, \ldots, k$.
- The statistical hypotheses in this case are as follows:

$$H_0 : \beta_i = 0$$

- If the assumptions of the model are met then the sampling distribution of $\hat{\beta}$ is $N(\mu = \beta, \hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1})$. [a]
- The test statistic in this case is Student's t:

$$t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} = \frac{\hat{\beta}_i}{s^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{ii}} \sim t_{n-(k+1)}$$

- A confidence interval for the parameter, β_i, can be obtained as follows:

$$\hat{\beta}_i \pm t_{\nu,\alpha/2} \times s_{\hat{\beta}_i}$$

[a] Can be proved that $\hat{\sigma}^2_{\hat{\beta}_j} = s^2 / n s_j^2 (1 - R_j^2)$

Multiple Linear Regression Models

Predictive Capacity of the Model

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The variability of the random error σ^2 is related to the utility of the fitted model. The greater the variability of the random error, the poorer the fitted model in terms of the estimation of the parameters and the prediction.
- σ^2 is unknown and the best estimator is s^2 which can be obtained as follows:

$$\hat{\sigma^2} = s^2 = \frac{SSE}{n-(k+1)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}$$

- Some statistical packages provides with the so-called Root of Mean Square Error: RMSE = $s = \sqrt{s^2}$.
- Recall that the utility of the model implies the extent in which we can adequately predict the response by means of the model.
- Separate tests for the predictors will increase the Type I error rate. Therefore, we need a global test:

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 &: \text{At least one of the parameters is not equal to zero} \end{aligned}$$

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- The test statistic is the following:

$$F = \frac{(SS_y - SSE)/k}{SSE/(n-(k+1))} = \frac{R^2/k}{(1-R^2)/(n-(k+1))} = \frac{MS_{\text{Model}}}{MS_{\text{Error}}}$$

- Under the null hypothesis: $F \sim F_{k;n-(k+1)}$
- Statistical significance:

$$p - value = P[F_{\nu_1,\nu_2} \geq F_{Obs} \mid H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0]$$

- Recall that by means of the coefficient of determination we quantify the capacity of the model to predict the response.

$$R^2 = 1 - \frac{SSE}{SS_{yy}}$$

- It ranges from 0 to 1 though is usually expressed as a percentage.
- R^2 can be interpreted as the percentage of the response variable that can be explained by the model.
- In the case of the multiple regression model R^2 can be misleading. A better indicator:

$$R^2_{adj} = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-(k+1))}$$

Multiple Linear Regression Models

Multicollinearity Problem

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- Multicollinearity is due to the correlation between two or more predictors.
- Problems arise when high correlation exists between the predictors.
- Recall that the OLS estimates of the coefficients of a multiple linear regression are: $\hat{\beta} = (X'X)^{-1} X'Y$.
- When two or more predictors are highly correlated, the previous inverse cannot be solved.
- In presence of multicollinearity coefficients cannot be interpreted due to unreliable estimates and statistical tests are biased.
- As stated previously, in presence of multicollinearity, inflated standard errors of the coefficients cause non-significant partial tests.
- The relationship between the variance of the coefficient and the variance of the error term is:

$$s^2_{\beta_i} = s^2 \left(\frac{1}{1 - R^2_i} \right)$$

where R^2_i is the coefficient of determination of the $i - th$ predictor and the remaining predictors. Therefore, when R^2_i is large, the standard error of the coefficient will be large.

Multiple Linear Regression Models

Multicollinearity Problem

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- The variance inflation factor (VIF) is an indicator of multicollinearity and can be obtained as follows:

$$VIF = \frac{1}{1 - R_i^2}$$

- We cannot usually state that there's no multicollinearity given that frequently there's some degree of association between predictors.
- If there's no multicollinearity at all, VIF will be equal to 1.
- Multicollinearity can be concluded as VIF increases its value.
- $1 < VIF < 5$ indicates a low level of multicollinearity.
- $5 < VIF < 10$ indicates a medium level of multicollinearity.
- $10 < VIF$ indicates a high level of multicollinearity, that is, a problematic one.

Multiple Linear Regression

Example

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- Social prestige as a function of income and women percentage.

	Dependent variable:
	prestige
income	0.003^{***}
	(0.0003)
women	0.133^{***}
	(0.040)
Constant	20.327^{***}
	(2.996)
Observations	102
R^2	0.559
Adjusted R^2	0.550
Residual Std. Error	11.537 (df = 99)
F Statistic	62.812^{***} (df = 2; 99)
Note:	*p<0.1; **p<0.05; ***p<0.01

Multiple Linear Regression

Example

Simple Linear Regression

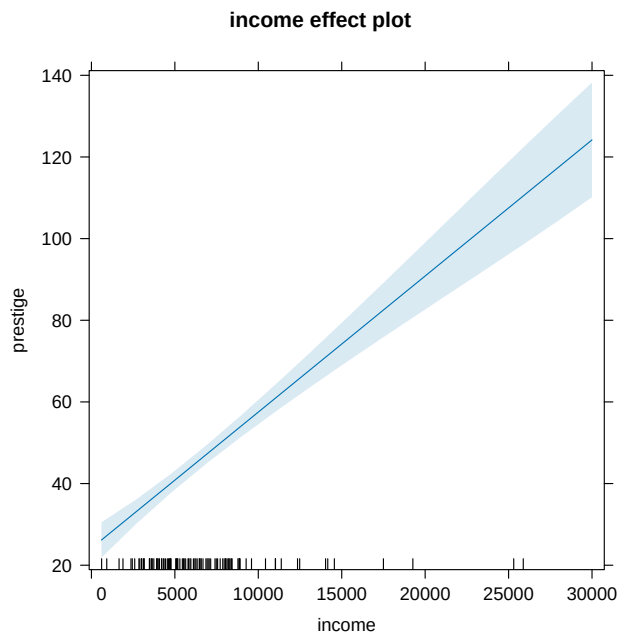
Models diagnostics

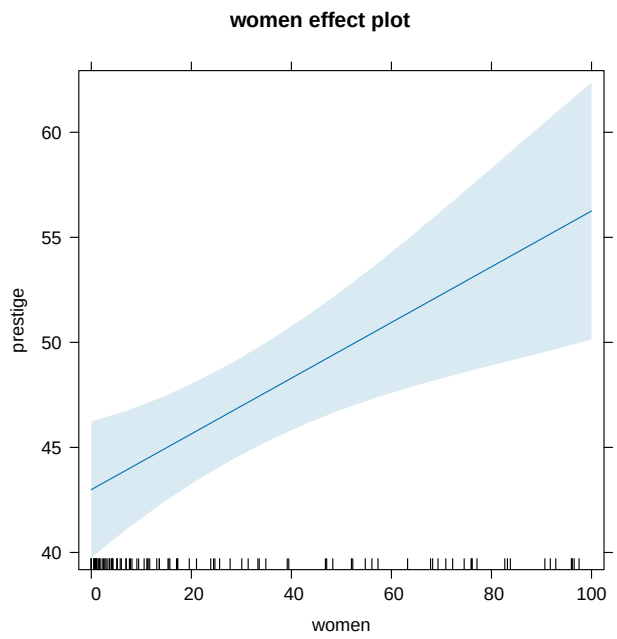
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Social prestige as a function of income and women percentage.





Qualitative Predictors

Definitions and notation

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- One of the requirements of the regression models seen so far is that both the response variable and the predictors are quantitative.
- How can we build a model that includes a qualitative predictor, that is, a categorical variable?
- To answer this question we need to introduce a general analytical framework: the General Linear Model. The basic notation of the General Linear Model is:

$$\begin{aligned} Y &= \mathbf{X}\beta + \epsilon \\ \epsilon &\sim \text{Normal}\left(0, \sigma^2 \mathbf{I}_n\right) \\ Y|X &\sim \text{Normal}\left(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n\right) \end{aligned}$$

$$Y_{n\times 1} = X_{n\times(k+1)}\beta_{(k+1)\times 1} + \epsilon_{n\times 1}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Qualitative Predictors

Definitions and notation

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Suppose that we want to include vendor's gender, x_3, as another predictor in the previous model.
- A way to express the different categories of the predictor in the model is required.
- Thus, different levels are assigned to the categories. The way to assign different values to the categories is by means of the so-called *dummy variable*.
- Gender can be coded as follows:

$$x_3 = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

- Then, the model will be the following:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Qualitative Predictors

Definitions and notation

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- The design matrix (X) in this case, will include a $0-1$ column for this new predictor (corresponding to the dummy variable):

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 32 & 5.5 & 0 \\ 1 & 23 & 10 & 1 \\ \vdots & \vdots & \vdots & \\ 1 & 45 & 8.5 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- How is the coefficient of the qualitative predictor interpreted?

$$\begin{aligned} E(y|x = \text{Female}) &= E(y|x = 0) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ E(y|x = \text{Male}) &= E(y|x = 1) = \beta_0 + \beta_3 + \beta_1 x_1 + \beta_2 x_2 \end{aligned}$$

Qualitative Predictors

Definitions and notation

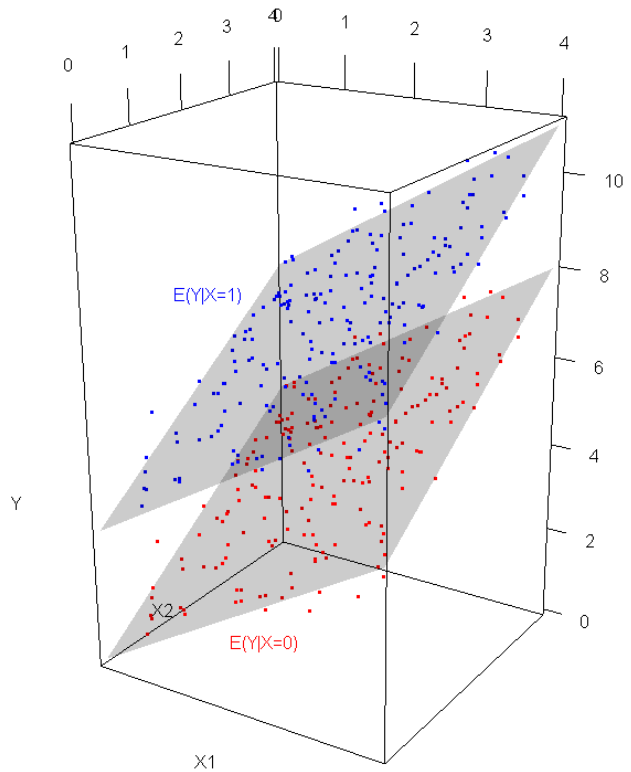
Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



Qualitative Predictors

How to code qualitative predictors

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- Suppose we have a qualitative predictor with 4 categories. One could think that a good way of including it in the model is as follows:

Original variable	D1	D2	D3	D4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1

- Nevertheless, when including these variables in the design matrix jointly with the column of the intercept (first column of 1's in matrix $\mathbf{X}$), the problem cannot be solved ($(\mathbf{X}'\mathbf{X})^{-1}$ doesn't exist).
- A qualitative predictor with k categories can be included in a regression model by means of $k-1$ dummy variables.

Qualitative Predictors

How to code qualitative predictors

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- Different **coding schemes** for categorical/ordinal predictors can be used depending on the hypotheses of interest and the type of predictors.
- Broadly speaking, we need to impose a restriction to the model. In this type of contrast we force the model to accomplish: $C_i = 0$, for $i = 1, \ldots, k-1$ in the reference category.

Original variable	D1	D2	D3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

- In this contrast we assume that $\mu = \mu_1$ and $\beta_i = \mu_i - \mu_1$.
- In R, the reference is the first category. In SPSS and SAS is the last one.

Qualitative Predictors

Interaction between a qualitative and a quantitative predictor

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

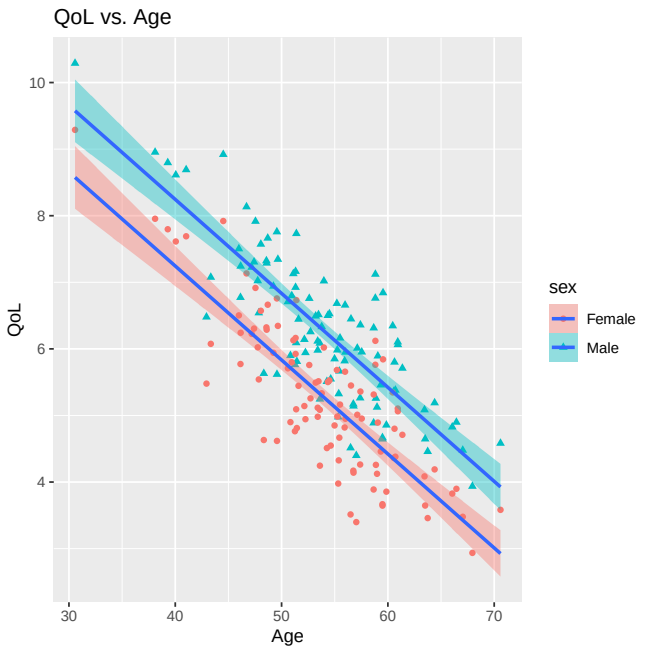
Polynomial Regression Models

- Let's focus on a model that predicts QoL as a linear function of gender, age of caregivers and their interaction.The model's equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- If the cofficient regarding the qualitative predictor ($X1$) leads to a significant result, we can assume a parallel lines model.
- Thus, the two regression lines only differ in the intercept. That is, the Wald test would lead us to reject at least:

$$H_0 : \beta_1 = 0$$



Qualitative Predictors

Interaction between a qualitative and a quantitative predictor

Simple Linear Regression

Models diagnostics

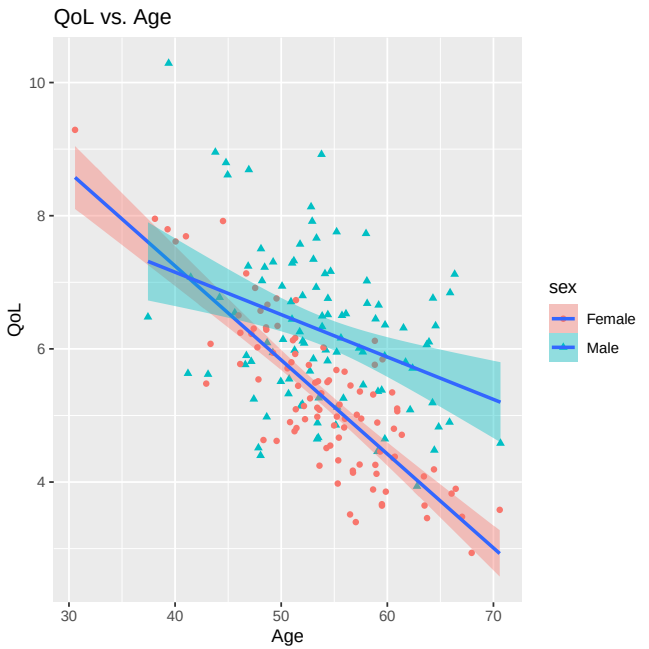
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- If the cofficient regarding the interaction between the qualitative and the quantitative predictors leads to a significant result we can assume a two lines model.
- Thus, the two regression lines differ in the slopes. That is, the Wald test would lead us to reject:

$$H_0 : \beta_3 = 0$$



Qualitative Predictors

Interaction between a qualitative and a quantitative predictor

Simple Linear Regression

Models diagnostics

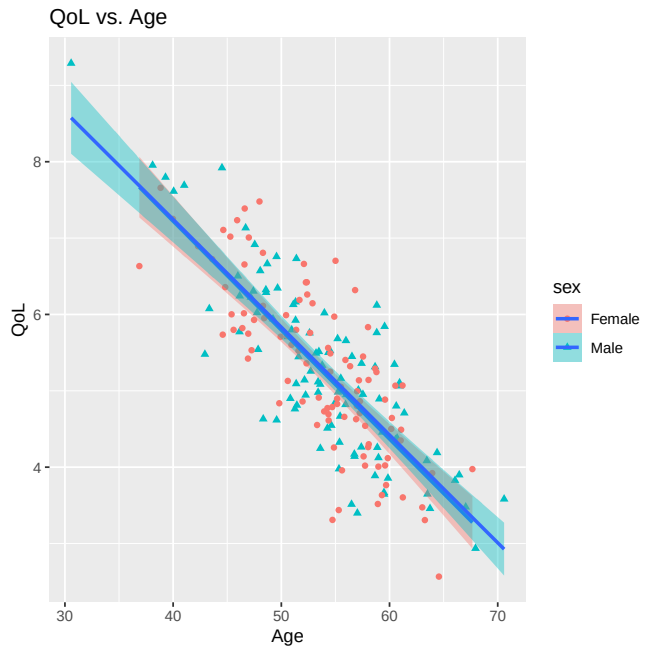
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- If the cofficient related to the main effect of the quantitative predictor is the only one that leads to a significant result then we can assume a coincident lines model.
- Thus, we can only conclude that the slope for the quantitative predictor is significantly different from 0. That is, the Wald test would lead us to reject:

$$H_0 : \beta_2 = 0$$



Polynomial Regression

Introduction

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

- So far we have assumed linearity regarding the relationship between the parameters of the model as well as concerning the variables.
- Nevertheless, we can build models with non-linear relationships between the variables, that is, a model that takes into account a curvature in the relationship between the response and the quantitative predictors. For instance, a quadratic model.
- Polynomial models are a specific kind of multiple regression models.
- Therefore, they can be estimated by means of Ordinary Least Squares and the assumptions of the linear regression model seen on previous sessions are also applied here.
- The general notation of a k-th order polynomial model is:

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

- In order to obtain an interpretable predictive/explicative model we need to choose the most parsimonious one.
- A trivial solution is to fit a polynomial model of order $n - 1$. That guarantees a perfect fit but as a result we obtain a useless model.

Polynomial Regression

Introduction

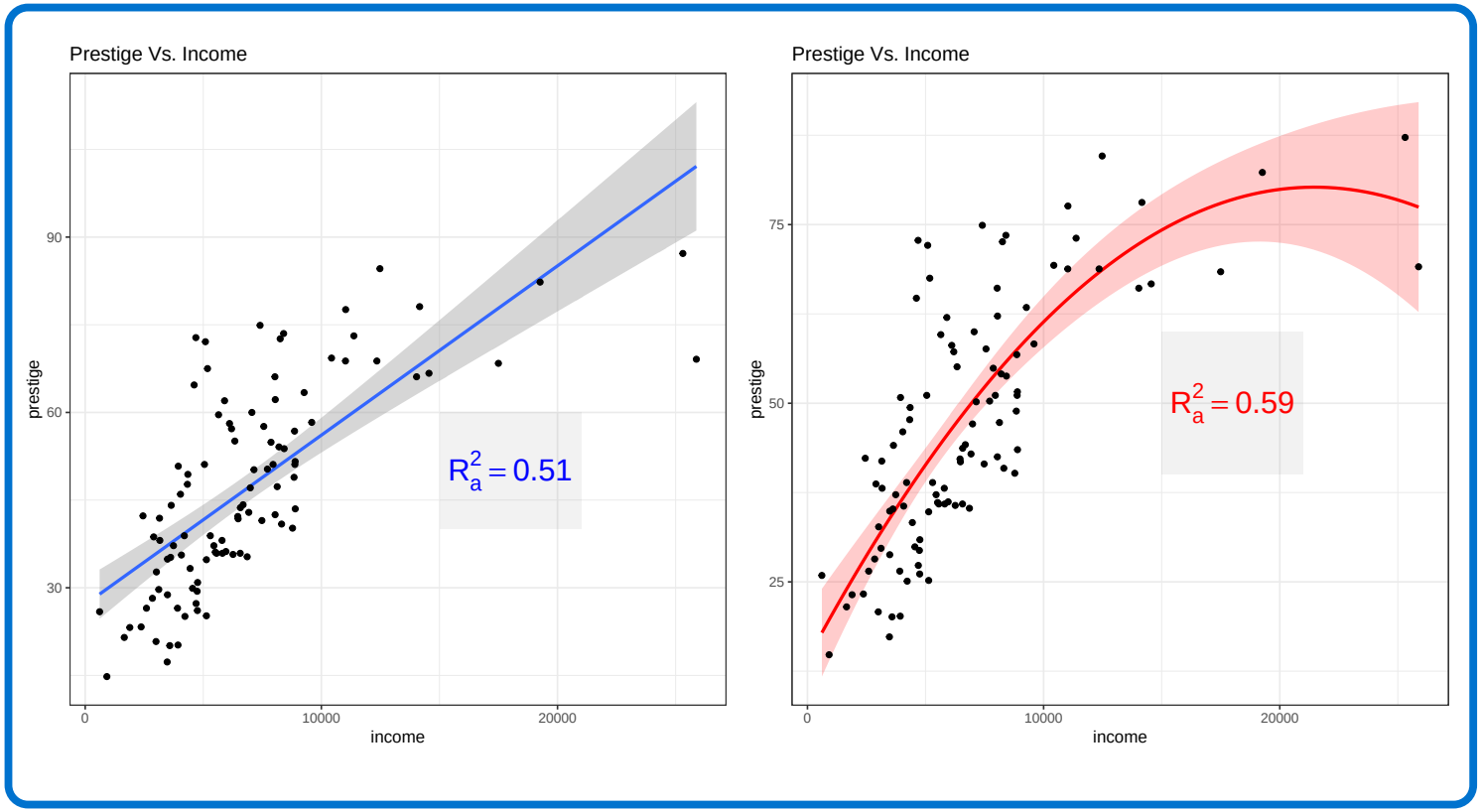
Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



Polynomial Regression

Introduction

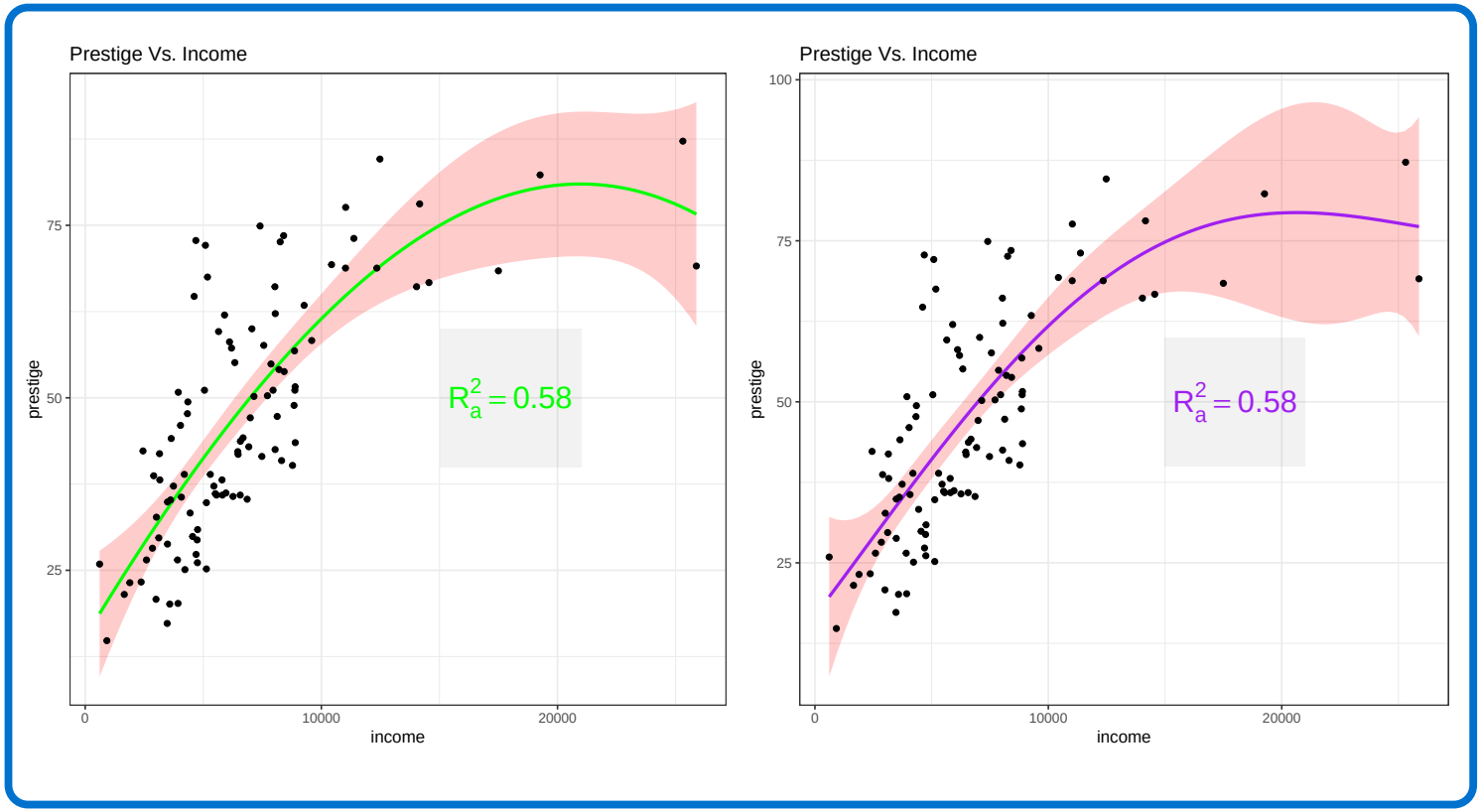
Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models



Interaction models

2nd order models

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- Including interactions between quantitative predictors allows us to account for certain types of non-linear relationships.
- The general notation of a *2*-nd order model is:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- β_0 is the intercept, that is, $E(y)$ when $x = 0$.
- β_1 and β_2 correspond to the main effects of the two predictors.
- β_3 is the interaction term.

Interaction models

2nd order models

Simple Linear Regression

Models diagnostics

Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

- As for the interpretation of the in interaction term, it should be thought to the element that causes the 'regression plane' to bend.
- $\beta_1 + \beta_3 X_2$: is the change in $E(y)$ for a unit increase in X_1 when X_2 is held fixed.
- $\beta_2 + \beta_3 X_1$: is the change in $E(y)$ for a unit increase in X_2 when X_1 is held fixed.
- The interaction terms can be classified as sinergistic (same signs for main effects and interaction term) or antagonistic (opposite signs).

Interaction models

Example

Simple Linear Regression
Models diagnostics
Multiple Linear Regression Models
Qualitative Predictors
Polynomial Regression Models

Income and women interaction: 1st degree polynomial (linear)

	Dependent variable:
	prestige
income	0.003***
	(0.0003)
women	−0.165**
	(0.075)
income:women	0.0001***
	(0.00002)
Constant	23.923***
	(2.847)
Observations	102
R^2	0.636
Adjusted R^2	0.625
Residual Std. Error	10.535 (df = 98)
F Statistic	57.125*** (df = 3; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interaction models

Example

Simple Linear Regression

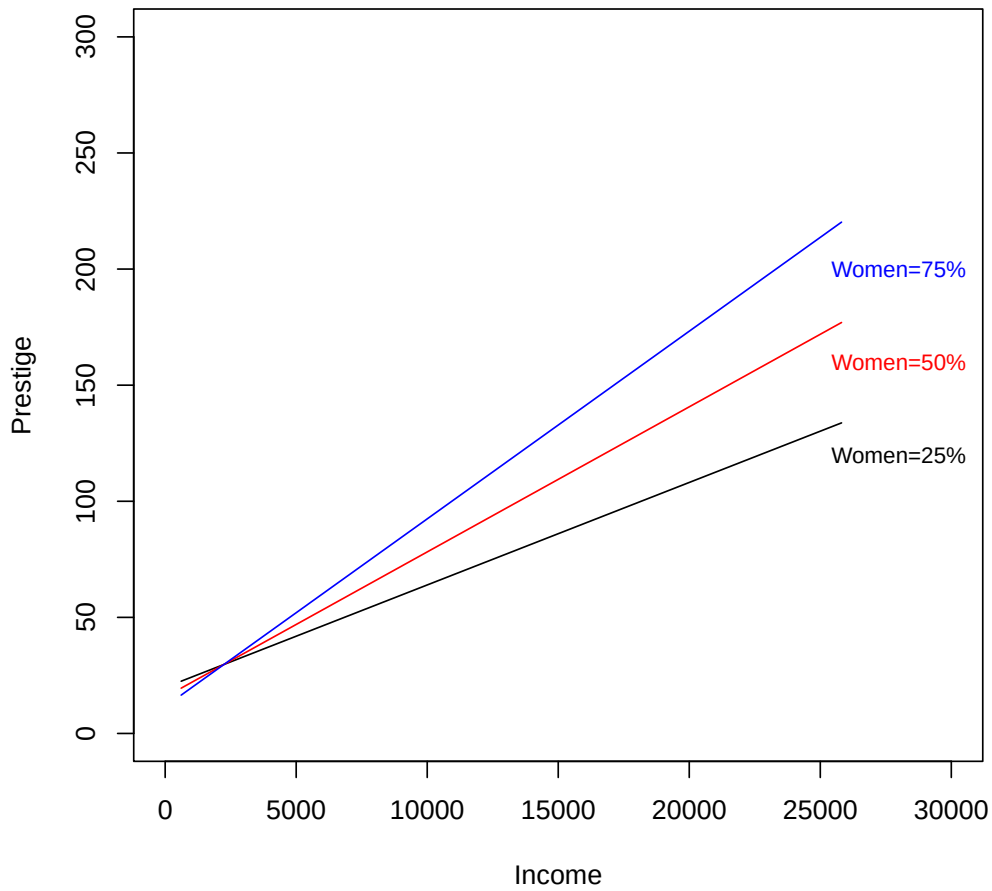
Models diagnostics

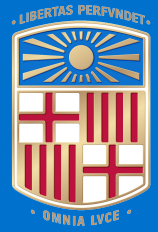
Multiple Linear Regression Models

Qualitative Predictors

Polynomial Regression Models

Income and women interaction: 1st degree polynomial (linear)





UNIVERSITAT DE BARCELONA

This work is licensed under the *Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License*.

Slides made with `knitr` and LaTeX