Section 7: Regression Analysis with R

STA 141A – Fundamentals of Statistical Data Science (UC Davis, Spring 2026)

Instructor: Akira Horiguchi

Overview

Based on Chapter 3 of ISL book James et al. (2021). For more R code examples, see R Markdown files in book website.

- I think you all have used the textbook Applied Linear Statistical Models by Kutner et al. (5e) in either STA 108 (first half of book) or 106 (second half of book). Half of a book can go into much more detail than contained in Ch3 of ISLR.
- The material in this slide deck is more to help with your project. The second midterm exam will not test on this much, if at all. (I prefer to test on things you haven't seen in the prerequisite courses.)
- StatQuest: Linear Regression, Clearly Explained!!! (27:26)
- StatQuest: Multiple Regression, Clearly Explained!!! (5:24)
- For ggplots in this slide deck, assume we are using `theme_minimal()` so that the slides don't have to show that line every time it is used.

Regression

Linear Regression

Idea of polynomial regression

Regression

Linear Regression

Idea of polynomial regression

An example – 1

Consider the data set in `Advertising.csv` consisting of the sales of a product in 200 different markets, with advertising budgets for the product in each of those markets for three different media: TV, Radio, Newspaper.

```
adv <- read.csv("Advertising.csv", row.names="X")
str(adv)
```

```
> str(adv)
'data.frame':   200 obs. of  4 variables:
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

An example – 2

We want to investigate the relationship between `Sales` and the total budget spent for advertisement on `TV`, `Radio`, and `Newspaper`.

- Then, we sum row-wise, but exclude the last column (which is the 4th column after we deleted the 1st column).

```
adv$Budget <- rowSums(adv[, -4])
str(adv)
```

```
> str(adv)
'data.frame':   200 obs. of  5 variables:
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
 $ Budget   : num  337 129 132 251 250 ...
```

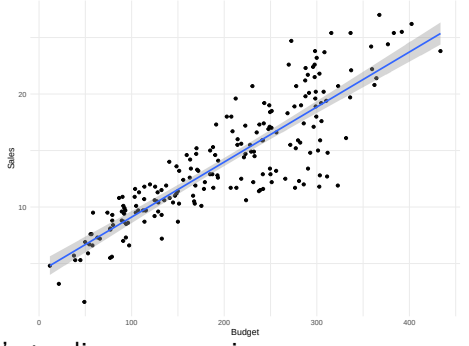
An example – 3

Reasonable research questions for this data set:

- Is there a relationship between `Budget` and `Sales`?
- If there is a relationship, is it linear?
- How strong is the relationship between `Budget` and `Sales`?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?

An example – 4

```
ggplot(adv, aes(Budget, Sales)) +
  geom_point() +
  geom_smooth(method="lm")
```



Relationship seems linear. Let's try linear regression.

Regression

Linear Regression

Idea of polynomial regression

Linear regression – motivation

Recall our general regression model:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon \tag{1}$$

where the error term ε is a catch-all for what is missed by this model.

- It is difficult to estimate f without any restrictions on what type of estimates $\hat{f}$ we allow.
- (More sophisticated math, like *functional analysis*, would allow more flexible regression methods.)
- *Linear regression* simplifies the task of estimating an arbitrary function f in (1) to the much easier task of estimating $p + 1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{2}$$

- This simplification helps with inference as well; we can easily interpret $\beta_0, \beta_1, \ldots, \beta_p$.

Linear regression – matrix representation

- For n observations $(y_1, \mathbf{X}_1), (y_2, \mathbf{X}_2), \ldots, (y_n, \mathbf{X}_n)$ we usually write

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \text{for } i = 1, 2, \ldots, n. \tag{3}$$

  where the errors ε_i follow a distribution with mean zero and variance σ^2.

- However, it is often more convenient to use the following matrix representation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \Longleftrightarrow \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

  where the $n \times (p+1)$ matrix $\mathbf{X}$ is known as the *design matrix*.

- Among many advantages, this matrix formulation enables compact representations.
- For a stats/DS career, I encourage you to learn as much linear algebra as you can.

Ordinary Least Squares (OLS)

A "good" estimator $\hat{f}$ for the relationship f in (1) is one that produces small residuals.

- A *residual* is the difference between a response value y at x and its predicted value $\hat{f}(x)$.
- The residual for the ith observed data point is $y_i - \hat{f}(x_i)$ for $i = 1, \ldots, n$.
- Recall the *residual sum of squares (RSS)*:

$$\sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 \qquad \text{or, more compactly,} \qquad \|\mathbf{y} - \hat{f}(\mathbf{X})\|^2 \tag{5}$$

where $\|\mathbf{z}\|$ is just the usual Euclidean length of vector $\mathbf{z}$,
i.e., for any $\mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{R}^n$ holds $\|\mathbf{z}\|^2 = z_1^2 + \cdots + z_n^2$.

In linear regression, we have $\hat{f}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, so $\hat{f}$ is parameterized by $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$.

- The *Ordinary Least Squares (OLS) estimator* for $\boldsymbol{\beta}$ is the vector that minimizes the RSS (5):

$$\hat{\boldsymbol{\beta}}_{OLS} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{6}$$

$$= \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2. \tag{7}$$

- It can be shown that $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ (if the inverse $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists).

Ordinary Least Squares (OLS) – geometry

The OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} \in \mathbb{R}^{p+1}$ produces:

- $p = 0$: sample mean of the $y_1, \ldots, y_n$ (best estimate without predictor info)
- $p = 1$: "line of best fit"
- $p = 2$: "plane of best fit"
- $p \geq 3$: "hyperplane of best fit"

Residual is vertical displacement between point and line/plane/hyperplane:

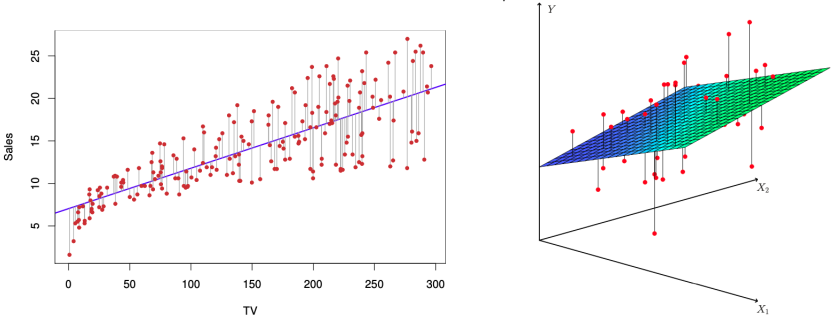


Figure 1: Image by James et al. (2021). The line (left) or plane (right) is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the line/plane.

(Linear) Regression – Questions of interest

Questions apply to regression generally, but answers might be specific to linear regression.

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response Y?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

1. Is there a relationship between the response and predictors?

Are all regression coefficients equal to zero?

- What would this imply about the relationship?
- One can use the hypothesis test

$$H_0 : \beta_1 = \beta_2 = \cdots \beta_p = 0 \qquad \text{vs.} \qquad H_a : \beta_j \neq 0 \text{ for at least one } j, \tag{8}$$

for which the following F-statistic is needed:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \geq 1, \tag{9}$$

where we define *total sum of squares (TSS)* and *residual sum of squares (RSS)*

$$TSS := \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = RSS_0 \qquad \text{and} \qquad RSS := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{10}$$

  - *TSS* measures the variability in the response values.
  - *RSS* measures the variability after performing the regression.
- If linear model assumptions are correct, then one can show that $\mathbb{E}[\text{denominator of } (9)] = \sigma^2$.
- If also H_0 is true, then one can show that $\mathbb{E}[\text{numerator of } (9)] = \sigma^2$.
  In this case, we expect $F \approx 1$.
- If H_a is true, we might expect $F > 1$.

2. Deciding on important variables

We might test whether one specific regression coefficient is zero or not.

- For a specific $j = 1, \ldots, p$, one can use the hypothesis test

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_a : \beta_j \neq 0 \tag{11}$$

  for which the following t-statistic is needed:

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}. \tag{12}$$

- $SE(\hat{\beta}_j)$ is the *standard error of* $\hat{\beta}_j$.
- Under H_0 holds $t_j \sim t_{n-p-1}$, where t_{n-p-1} is the *Student's t-distribution*.

2. Deciding on important variables

Variable selection (or model selection): the task of determining which predictors are associated with the response.

- A model contains a subset of the predictors.
- For p predictors, there are 2^p possible models.
- Many ways to choose a model in linear regression. See Ch 6 of ISLR2 for more details (e.g. subset selection, lasso).

Outside linear regression, this is still an active research area!

3. Model fit

Potential problems: i) Non-linearity

- Linear regression assumes a linear relationship between the predictors and the response.
- Residual plots can be used to detect non-linearity: If no pattern is visible, linearity is a reasonable assumption, otherwise not.
- If there are non-linear associations, a simple approach is to check whether non-linear transformations of the predictors help, such as $\log(X)$, $\sqrt{X}$, or X^2.

3. Model fit

Potential problems: ii) Correlation of the error terms

- The errors are assumed to be uncorrelated.
- If the errors are correlated, the estimated standard errors tend to underestimate the true standard errors.
- Correlations frequently occur in the context of time series, where observations are analyzed over time, e.g. daily temperatures.

3. Model fit

Potential problems: iii) Non-constant variances of the error terms (*heteroskedasticity*)

- The errors are assumed to be *homoskedastic* (i.e., error variances are constant across observations).
- The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely on this assumption.
- One can identify *heteroskedasticity* from the presence of a funnel shape in the residual plot.
- One possible solution is to transform the response Y by using a concave function such as $\log(Y)$ or $\sqrt{Y}$.

3. Model fit

Potential problems: iv) Outliers

- An *outlier* is a point which is far from the response value predicted by the model.
- Outliers can arise for a variety of reasons, e.g., incorrect recording of an observation during data collection.
- Outliers might inflate the variance estimate, and other measures.

3. Model fit

Potential problems: v) Collinearity

- *Collinearity* refers to the situation in which ≥ 2 predictors are closely related, resulting in uncertainty in the coefficient estimates, and thus in the standard error for $\hat{\beta}_j$ to grow.
- Collinearity can also be present between more than two variables (*multicollinearity*), even if no pair of variables are closely related.
- The *variance inflation factor* (VIF) quantifies the severity of (multi)collinearity: It measures how much the variance of an estimated regression coefficient is increased due to collinearity.
- VIF for each variable can be computed by

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R_j^2},$$

  where R_j^2 is the R^2 from a regression of X_j onto all other predictors.
- As a rule of thumb, VIF exceeding 5 implies a large amount of collinearity, then we should ...
  1. ... drop the jth predictor;
  2. ... or combine the jth with other collinear predictors together into a single predictor.

3. Model fit: Estimating the variance

Consider the linear model with $n > p + 1$ where the columns of the $n \times (p + 1)$ covariance matrix $\mathbf{X}$ are *linearly independent* (i.e., cannot write a column as a linear combination of the others).

- The OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ for $\boldsymbol{\beta}$ is the *BLUE* (Best Linear Unbiased Estimator) if $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_n$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, where $\sigma^2 > 0$ is the error variance, and $\mathbf{I}_n$ is the $n \times n$ identity matrix.
- Usually, the variance σ^2 is unknown and has to be estimated. An unbiased estimator for σ^2 is

$$\hat{\sigma}^2 \;:=\; \frac{1}{n-p-1} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 \tag{13}$$

- For $p = 0$ (no predictor info is used), notice that $\hat{\sigma}^2$ above reduces to the sample variance

$$s^2 \;:=\; \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 \tag{14}$$

for $y_1, \ldots, y_n$ with $\bar{y}_n := \frac{1}{n}\sum_{i=1}^{n} y_i$. The R function `var()` computes s^2.

3. Model fit: Estimating the standard deviation

- If $\hat{\sigma}$ is almost surely (i.e. with probability 1) non-constant ($\hat{\sigma}$ is almost surely constant if all X_i equal to $\overline{X}$ with prob. 1), *Jensen's inequality* gives

$$\mathrm{E}(\hat{\sigma}) < \sqrt{\mathrm{E}(\hat{\sigma}^2)} = \sqrt{\sigma^2} = \sigma. \tag{15}$$

  This means that $\hat{\sigma}$ is not an unbiased estimator for σ, even though $\hat{\sigma}^2$ is an unbiased estimator for σ^2.
- (Don't need to know Jensen's inequality for this class, but it might be useful for future stats classes/work.)

3. Model fit: R^2 – Example: The data set

We consider the linear model with $p = 1$ and analyze the advertising data set.

```
adv <- read.csv('Advertising.csv')
fit <- lm(Sales ~ Newspaper, data = adv)
summary(fit)
```

```
Call:
lm(formula = Sales ~ Newspaper, data = adv)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2272  -3.3873  -0.8392   3.5059  12.7751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,   Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

3. Model fit: R^2 – Example: Comparison

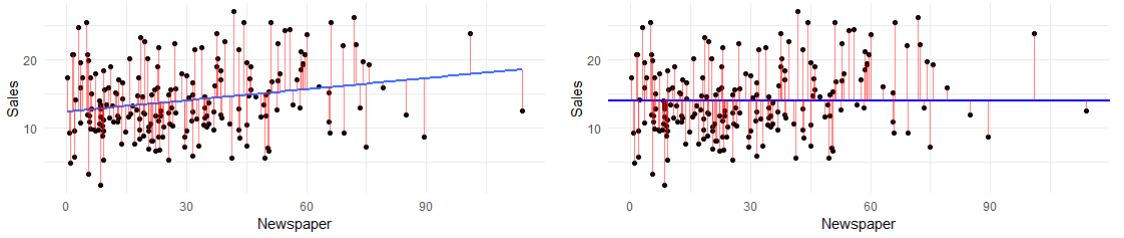


Figure 2: Fitted lines (blue) with residuals shown as red vertical lines. Right: line with zero slope.

Let's compare the SSE for this optimal fit to the SSE for the best zero-slope fit.

```
summary(fit)$df[2]   # (n - p - 1)
summary(fit)$df[2] * summary(fit)$sigma   # 198 * 5.092 = 1008
m <- mean(adv$Sales)   # 14
sum((adv$Sales - m)^2)   # 5417
```

Hence incorporating `Newspaper` shrunk the SSE by a factor of 5.

3. Model fit: R^2 – Definition and Interpretation

How much of the variability in Y is explained by $\mathbf{X}\hat{\boldsymbol{\beta}}$?

- Can measure goodness of the fit with the linear model using *coefficient of determination* R^2

$$R^2 := \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}. \tag{16}$$

  where TSS and RSS were defined in (10).
  - *total sum of squares (TSS)* measures the total variability in the response values.
  - *residual sum of squares (RSS)* measures the variability after performing the regression.
- In our example, we have $R^2 = 1 - \frac{1008}{5417} \approx 0.814$, meaning that approximately 81.4% of the variability has been explained by the regression.

Interpretation: by definition, R^2 is the proportion of the total variability minus the variability after the regression, in relation to the total variability. Hence, R^2 has values between 0 and 1.

- $R^2 \approx 1 \implies$ the regression explained a large proportion of the variability in the response.
- $R^2 \approx 0 \implies$ the regression did not explain much of the variability in the response. Maybe, because the linear model is wrong, or the inherent error variance σ^2 is large, or both.

3. Model fit: Adjusted R^2

Including more (not perfectly collinear) predictors into the model will always increase the explained variation. Hence we cannot use R^2 to select predictors.

- The *adjusted* R^2, denoted as $\overline{R}^2$, also measures how much variability have been explained by the regression, but can decrease if weak/unimportant predictors are added to the model. It is defined as

$$\overline{R}^2 := 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{17}$$

$$= 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}. \tag{18}$$

- $\overline{R}^2$ is smaller than R^2 if RSS $\neq 0$ and $p > 0$.
- What happens to $\overline{R}^2$ if we add unimportant predictors? Examine (18).
  - $\text{TSS}/(n-1)$ does not depend on predictors.
  - If RSS decreases slowly as p increases, then $\text{RSS}/(n-p-1)$ can increase as p increases.
- As smaller p make inference easier, one should choose p that maximizes $\overline{R}^2$.

3. Model fit: Residual plots: Idea

Residual plots show the fitted values $\hat{y}_i$ against the observed values y_i, or the predictor values x_i against the residuals $e_i := y_i - \hat{y}_i$.

- Residual plots are mainly useful for two things:
  1. To validate/reject the suggested model.
  2. To extract further information about the data.
- Residual plots can show the following behaviors, among others:
  1. The values in the residual plot are scattered around zero without a visible trend
     $\Longrightarrow$ model assumption is reasonable.
  2. The values in the residual plot exhibit a visible trend/pattern
     $\Longrightarrow$ model assumption is NOT reasonable.
  3. Scatter plot or residual plot exhibits unusual values being far away from most of data
     $\Longrightarrow$ outliers!
  4. The magnitudes of the measurement errors are not roughly constant across observations
     $\Longrightarrow$ heteroskedasticity (variance heterogeneity).

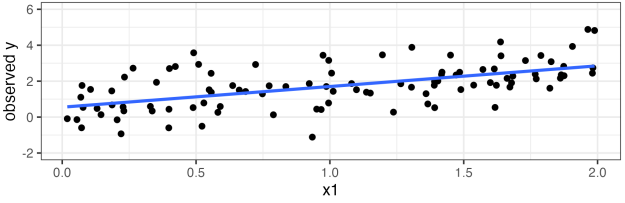
3. Model fit: Residual plots: Data set 1

We create the following data consisting of 100 rows.

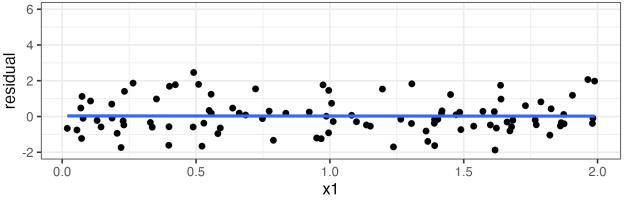
```
# Create data
df1 <- data.frame(x1=runif(n=100, min=0, max=2))
df1$yobs <- 0.1 + df1$x1 * 1.5 + rnorm(n=100, sd=1)

# Fit linear model to data and compute fitted values
fit1 <- lm(yobs ~ x1, data=df1)
df1$ypred <- fit1$coefficients[1] + df1$x1 * fit1$coefficients[2]

# Subtract fitted values from observed values
df1$residual <- df1$yobs - df1$ypred
```

3. Model fit: Residual plots: Data set 1

```r
# Plot line of best fit
ggplot(df1, aes(x1, yobs)) +
  geom_point() +
  geom_smooth(method='lm', se=F) +
  scale_y_continuous(limits=c(-2,6)) +
  labs(y='observed y')
```

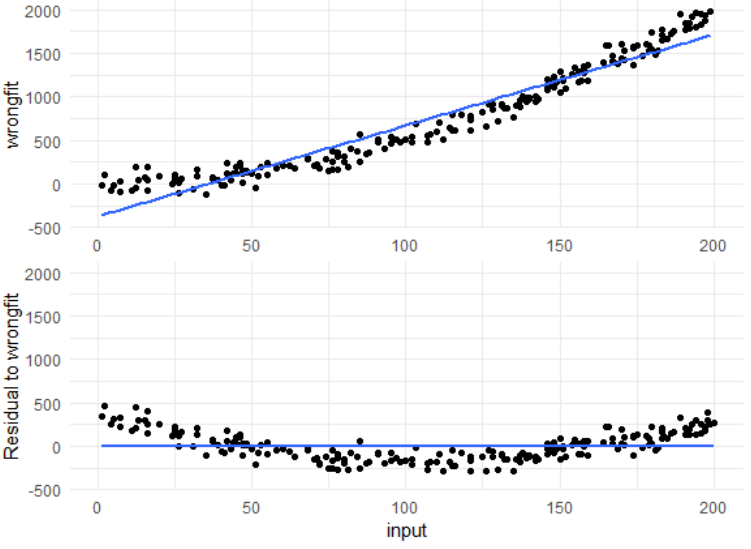



```r
# Plot residuals
ggplot(df1, aes(x1, residual)) +
  geom_point() +
  geom_smooth(method='lm', se=F) +
  scale_y_continuous(limits=c(-2,6))
```



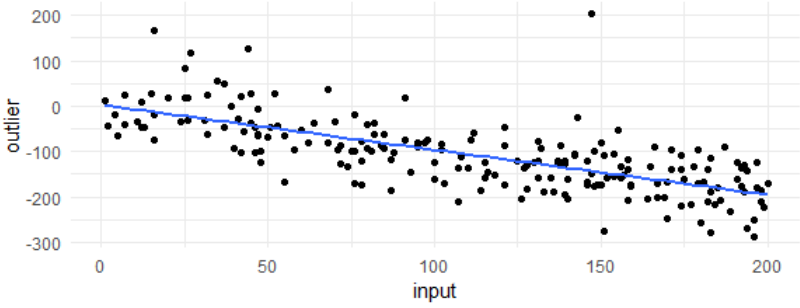
No visible pattern in residual plot $\Longrightarrow$ Proper fit of the data using a linear model

3. Model fit: Residual plots: Data set 2



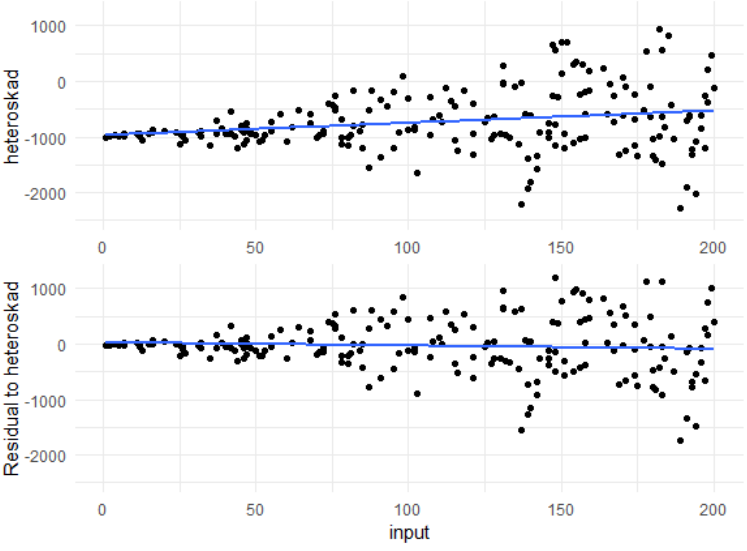
Visible pattern in the residual plot $\Longrightarrow$ The data are not properly fitted by the linear model. Maybe a quadratic relationship is reasonable.

3. Model fit: Residual plots: Data set 3



A very "unusual" value around $x = 150 \implies$ interpretable as an outlier.

3. Model fit: Residual plots: Data set 4



Error amplitudes increase as input increases $\implies$ Signal seems to be well-modeled as a linear function, but errors are heteroskedastic.

4. Predictions

With coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, it is straightforward to predict the response Y_{n+1} at a set of predictor values $\mathbf{X}_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})^\top \in \mathbb{R}^p$. (How?)

Three types of uncertainty associated with this prediction:

1. Inaccuracy in the coefficient estimates $\hat{\boldsymbol{\beta}}$ — quantify uncertainty using *confidence intervals*.
2. How well can the true model be captured by even the best linear model?
3. Inaccuracy in the prediction $\hat{Y}_{n+1}$ — quantify uncertainty using *prediction intervals*.
   - Width of prediction interval includes both model uncertainty and observation variance.

Regression

Linear Regression

Idea of polynomial regression

Polynomial regression extends the simple linear model...

...by also allowing sums of predictors raised by powers, thus "polynomial".

- In polynomial regression, the response Y is modelled depending on the predictor X_1 with a degree $d \in \mathbb{N}$ polynomial function:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \cdots + \beta_d X_1^d + \varepsilon, \tag{19}$$

- The degree d describes the flexibility of the model.
- (What does a polynomial of order $d = 2$ look like? Order $d = 3$? $d = 4$?)

Example 1: A non-linear function

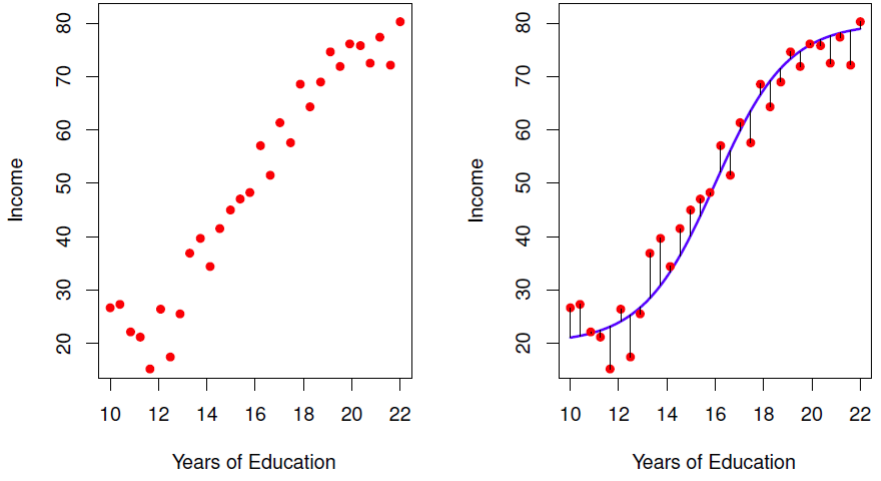


Figure 3: Image by James et al. (2021), based on the `Income` data set in `R`. The red dots are the observed values of `income` in tens of thousand dollars and `years of education` for 30 individuals.

Example 2: degree-4 polynomial

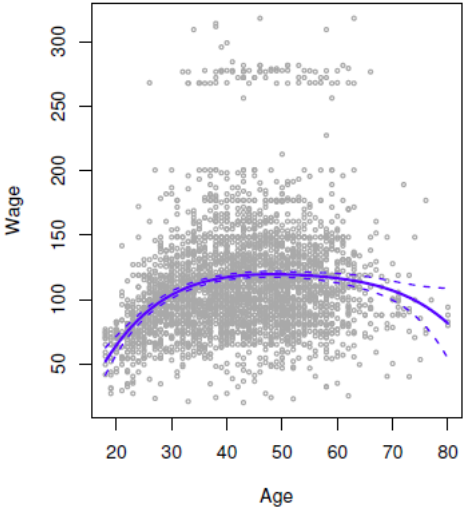


Figure 4: Image by James et al. (2021). The solid blue curve is a degree-4 polynomial of wage (in thousands of dollars) as a function of age, fit by least squares.

Example 3: Polynomial regression with two predictors

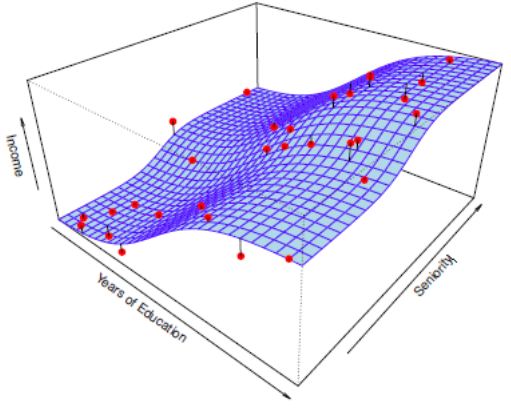


Figure 5: Image by James et al. (2021), based on the `Income` data set in `R`. The `income` is displayed as a function of `years of education` and `seniority`, where linearity does not seem appropriate. It might be reasonable to do polynomial regression with two predictors.