STA 141A – Fundamentals of Statistical Data Science

Department of Statistics; University of California, Davis

Instructor: Dr. Akira Horiguchi (ahoriguchi@ucdavis.edu)
A01 TA: Zhentao Li (ztlli@ucdavis.edu)
A02 TA: Zijie Tian (zijtian@ucdavis.edu)
A03 TA: Lingyou Pang (lyopang@ucdavis.edu)

Section 6: Regression Analysis with R

Spring 2025 (Mar 31 – Jun 05), MWF, 01:10 PM – 02:00 PM, Young 198

Overview

Based on Chapter 3 of ISL book James et al. (2021).

- For more R code examples, see R Markdown files in `https://www.statlearning.com/resources-second-edition`

Section 6: Regression Analysis with R

- Linear Regression
- Idea of polynomial regression

Section 6: Regression Analysis with R

Linear 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).

```
1 adv$Budget <- rowSums(adv[, -4])
2 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 ...
```

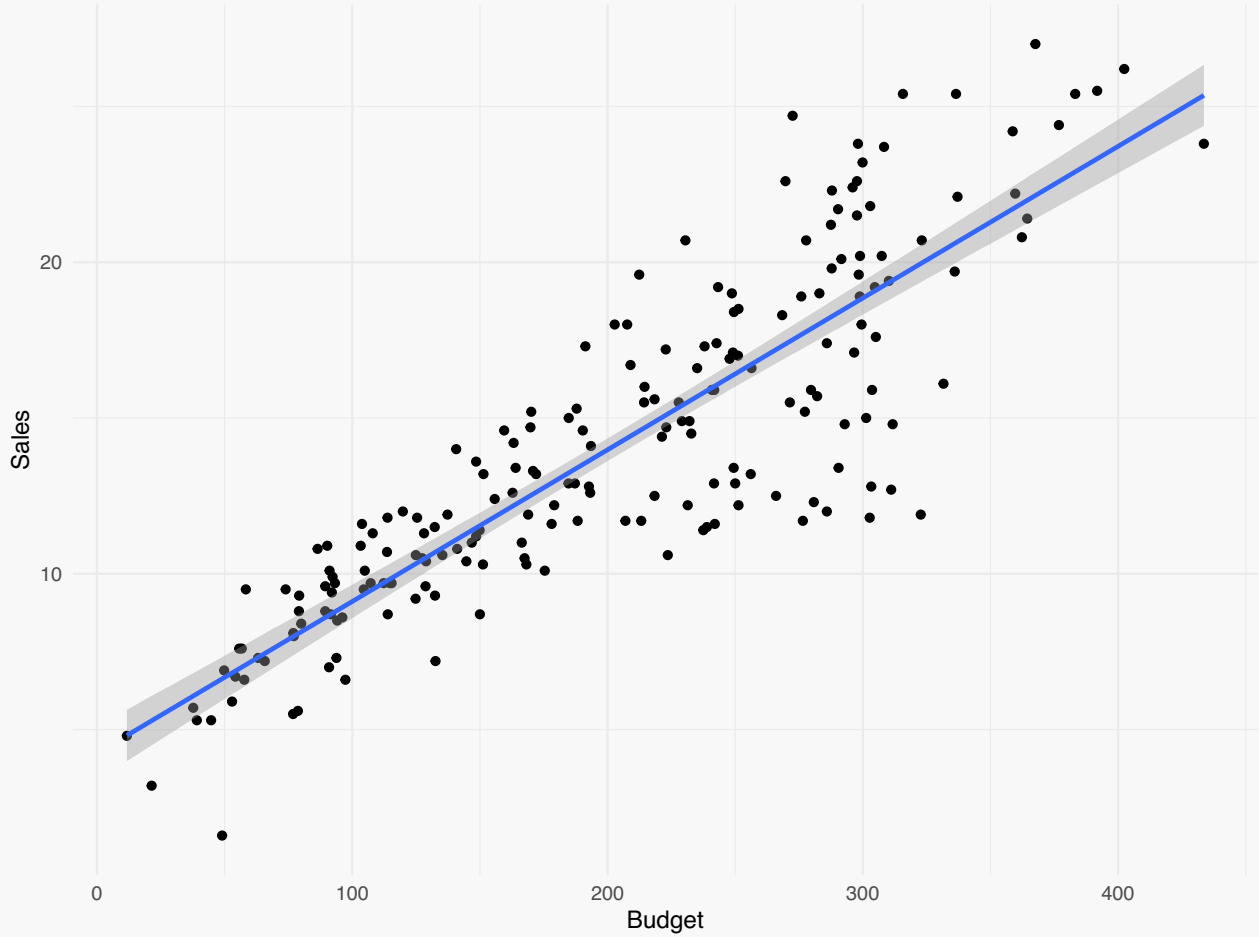
An example – 3

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") +
  theme_minimal()
```



Recall: want to estimate the relationship between Y and $X_1, \ldots, X_p$.

- Linear regression simplifies the task of estimating the "true" relationship to the task of estimating $p + 1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

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

- For n observations

$$\begin{aligned} (x_1, y_1) &= (x_{11}, x_{12}, \ldots, x_{1p}, y_1), \\ (x_2, y_2) &= (x_{21}, x_{22}, \ldots, x_{2p}, y_2), \\ &\vdots \\ (x_n, y_n) &= (x_{n1}, x_{n2}, \ldots, x_{np}, y_n), \end{aligned}$$

  there is the following matrix representation

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

$$\iff \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (3)$$

$$\iff \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4)$$

Ordinary Least Squares (OLS)

A "good" estimator for the regression parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ produces small residuals.

- A *residual* is the difference between a response and its predicted value.
- The ith *residual* is $y_i - \hat{y}_i$ for $i = 1, \ldots, n$.
- $\boldsymbol{\beta}$ can be estimated by using the *Ordinary Least Squares* (OLS) method.
- The *OLS estimator* for $\boldsymbol{\beta}$ is defined to be the vector that minimizes the residual sum of squares:

residual sum of squares

$$\hat{\boldsymbol{\beta}}_{OLS} := \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{5}$$

  where for any $\boldsymbol{z} = (z_1, \cdots, z_{p+1})^T \in \mathbb{R}^{p+1}$ holds $||\boldsymbol{z}||^2 = z_1^2 + \cdots + z_{p+1}^2$.
- It can be shown that $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ (if the inverse $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists).

Ordinary Least Squares (OLS) – Geometry

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

- $p = 0$: sample mean of the $y_1, \ldots, y_n$ (best estimate w/o predictor info)
- $p = 1$: "line of best fit"
- $p = 2$: "plane of best fit"
- $p \geq 3$: "hyperplane of best fit"

Given a line/plane/hyperplane of best fit:

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

Ordinary Least Squares (OLS) – Example line of best fit

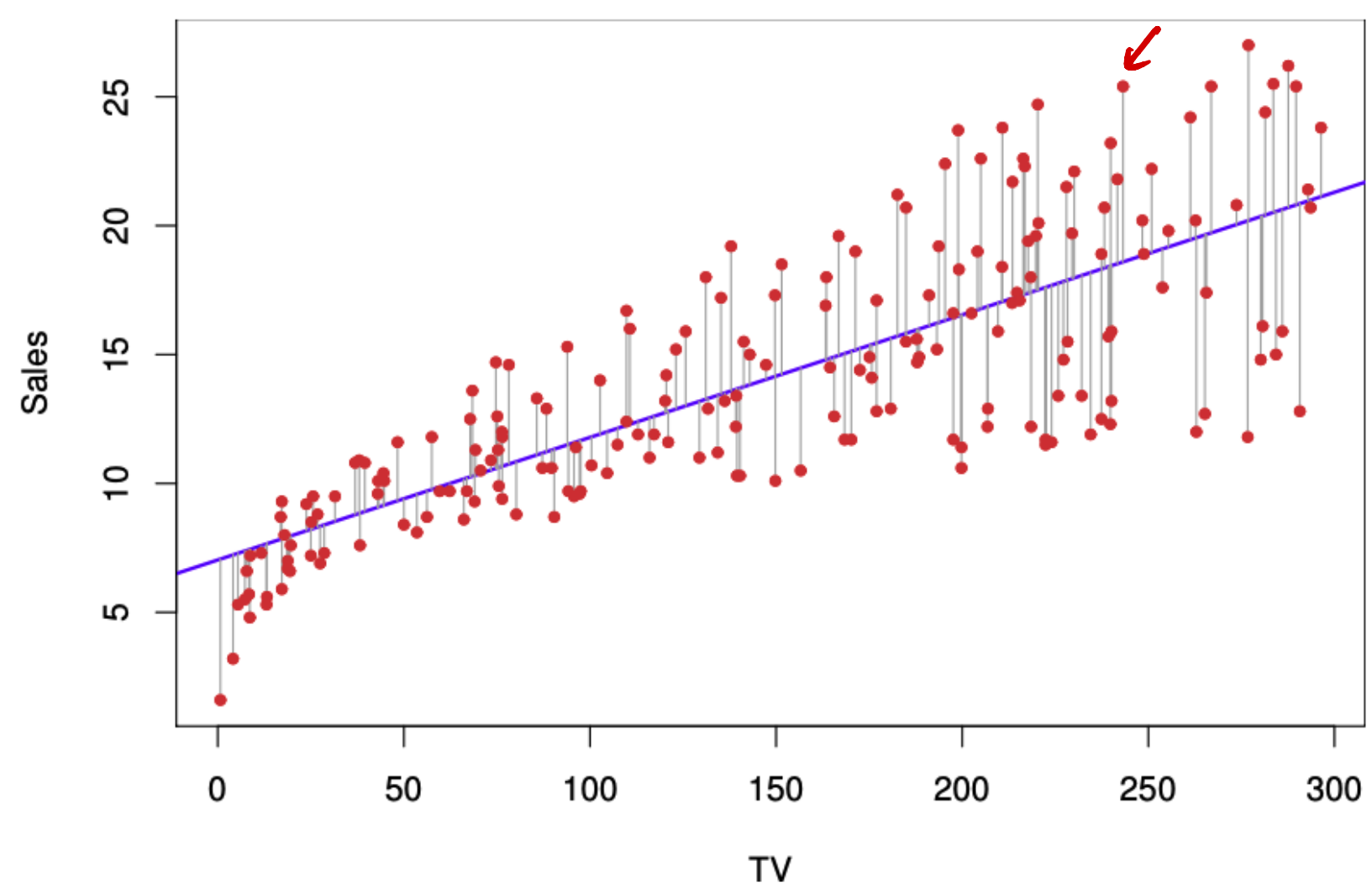


Figure 1: Image by James et al. (2021). "For the Advertising data, the least squares fit for the regression of sales onto TV is shown." The line is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the line.

Ordinary Least Squares (OLS) – Example plane of best fit

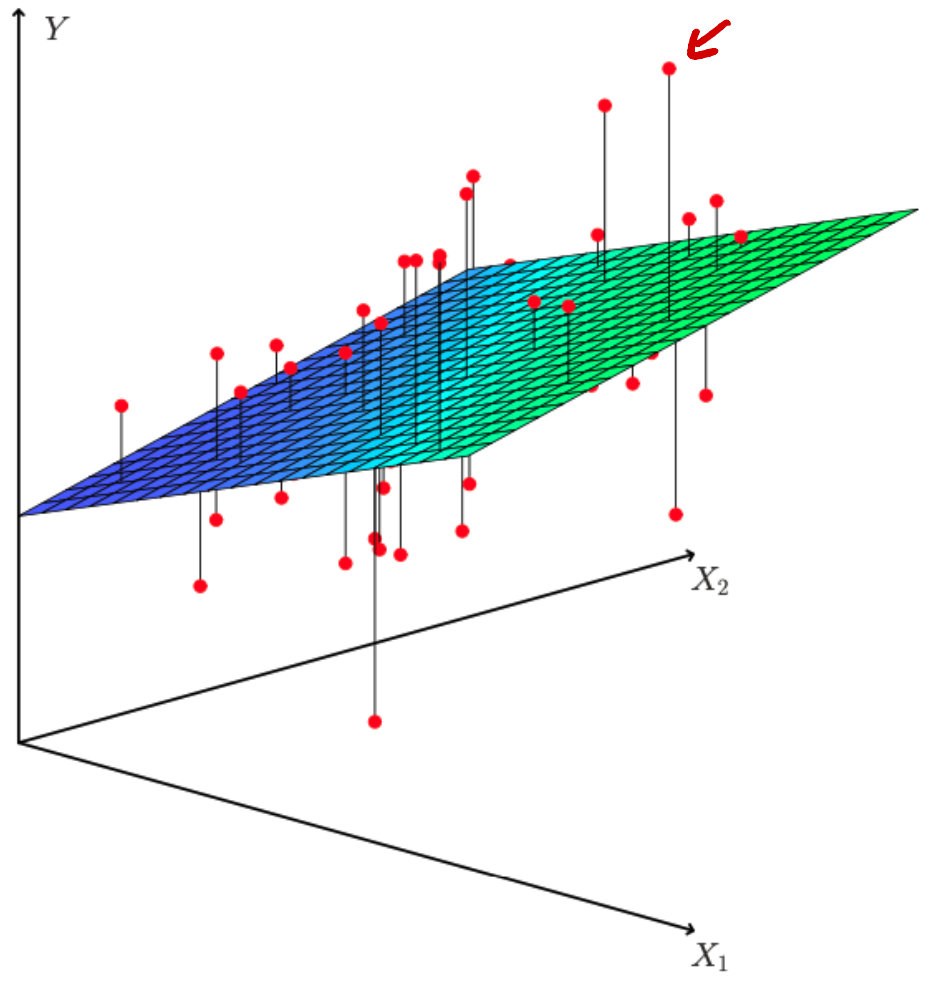


Figure 2: Image by James et al. (2021). "In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane."

End of May 2 lecture

Questions of interest:

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?

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

Are all regression coefficients equal to zero?

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

$$H_0 : \beta_1 = \beta_2 = \cdots \beta_p = 0 \quad \text{vs.} \quad H_a : \beta_j \neq 0 \text{ for at least one } j, \tag{6}$$

for which the following F-statistic is needed:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \geq 1. \tag{7}$$

- If H_0 is true, we might expect $F \approx 1$.
- If H_a is true, we might expect $F > 1$.

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

- For a specific $j = 0, \ldots, p$, one can use the hypothesis test

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_a : \beta_j \neq 0 \tag{8}$$

  for which the following t-statistic is needed:

$$t = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} . \tag{9}$$

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

Variable selection: the task of determining which predictors are associated with the response.

1. A model contains a subset of the predictors.
2. For p predictors, there are 2^p possible models.
3. Many ways to choose a model in linear regression.
4. Outside linear regression, this is still an active research area!

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.

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.

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

- The errors are assumed to be homoskedastic (meaning their 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 non-constant variances in the errors (*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 $\sqrt{Y}$.

Potential problems: iv) Outliers

- An *outlier* is a point which is far from the 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, the RSE, and other measures.

Potential problems: v) Collinearity

- *Collinearity* refers to the situation in which two or more 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 indicates 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.

Consider the linear model with $n > p + 1$, where the vectors of the predictors $X_1, \ldots, X_n \in \mathbb{R}^p$ are linearly independent.

- In the linear model the OLS estimator $\hat{\beta}_{OLS}$ for β is the *BLUE* (Best Linear Unbiased Estimator) if $\mathrm{E}(\varepsilon) = \mathbf{0}_n$ and $\mathrm{Var}(\varepsilon) = \sigma^2 \boldsymbol{I}_n$, where $\sigma^2 > 0$ is the variance, and $\boldsymbol{I}_n$ 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 given by

$$\hat{\sigma}^2 := \frac{1}{n-p-1} ||\boldsymbol{y} - \boldsymbol{X}\hat{\beta}||^2 \tag{10}$$

- For $p = 0$, we can use the sample variance s^2 in R (command (`var`)), which is for an independent sample $Y_1, \ldots, Y_n$ with $\overline{Y} := \frac{1}{n}\sum_{i=1}^n Y_i$ defined by

$$\hat{\sigma}^2 := \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{11}$$

Exercise: Show that $\hat{\sigma}^2$ is an unbiased estimator for the variance σ^2 in our model, i.e. show that $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

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

$$\mathrm{E}(\hat{\sigma}) < \sqrt{\mathrm{E}(\hat{\sigma}^2)} = \sqrt{\sigma^2} = \sigma. \tag{12}$$

  This means that $\hat{\sigma}$ is not an unbiased estimator for σ, although $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

Exercise: Illustrate the bias of $\hat{\sigma}$ for σ, that is $\hat{\sigma} - \sigma$:

1. Generate a random vector x of length $n = 10$ with i.i.d. $N(0, 4)$-distributed entries
2. Calculate the bias $\hat{\sigma} - \sigma$
3. Repeat the steps $nsim = 10,000$ times and report the average bias.

R^2 – Example: The data set

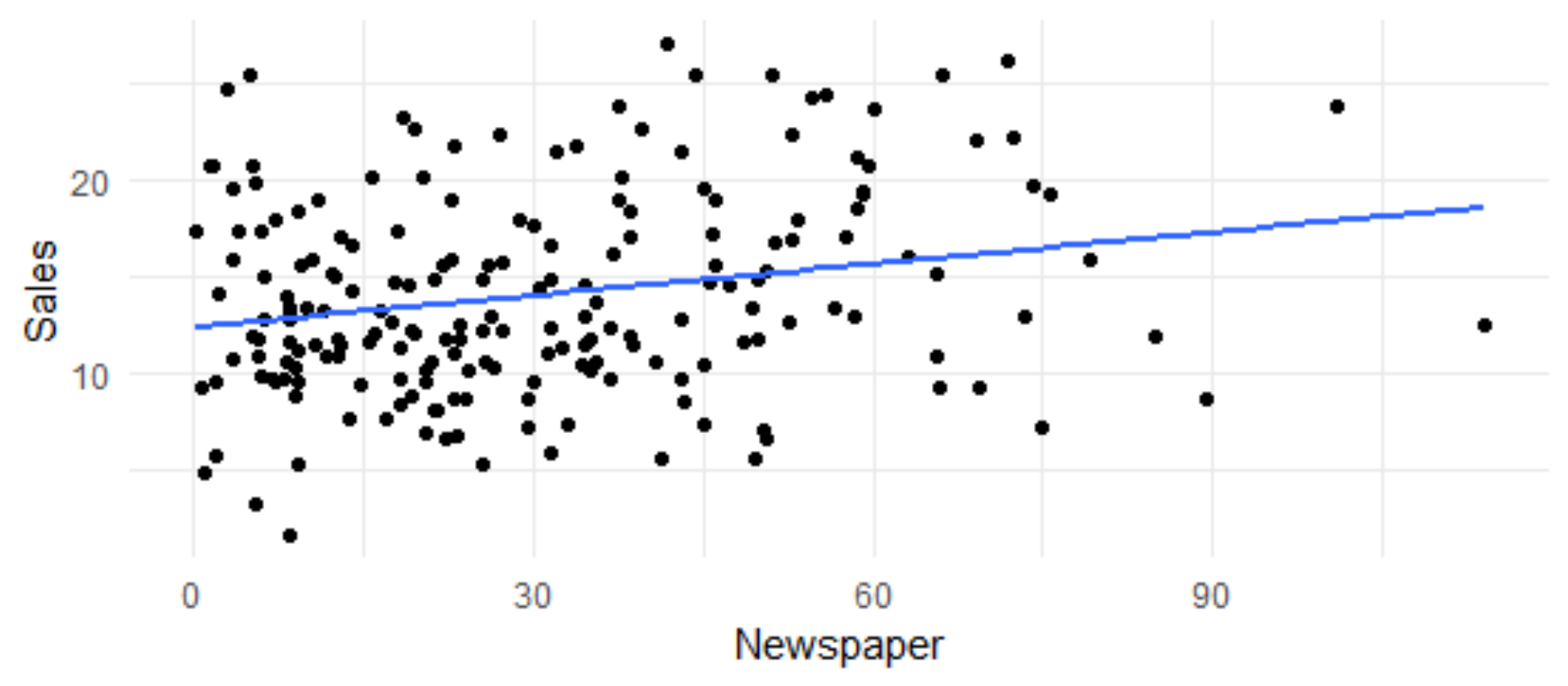
We consider the linear model with $p = 1$ and analyze the advertising data set.

```
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
```

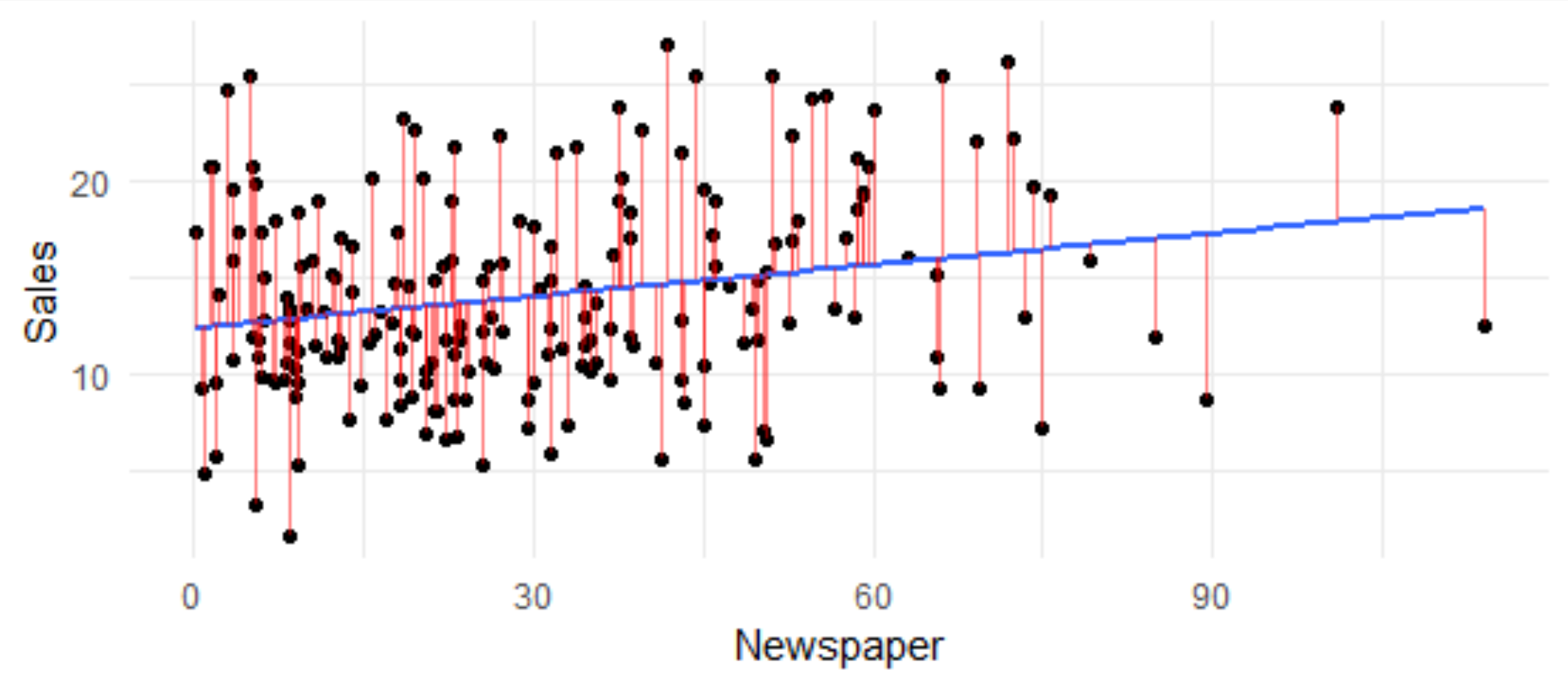
R^2 – Example: Regression line

```
ggplot(adv, aes(Newspaper, Sales)) +
    geom_point() +
    geom_smooth(method = "lm", se = F) +
    theme_minimal()
```



R^2 – Example: Regression line (errors visible)

```
fit <- lm(Sales ~ Newspaper, data = adv)
ggplot(adv, aes(Newspaper, Sales)) +
    geom_point() +
    geom_smooth(method = 'lm', se = F) +
    geom_segment(aes(xend=Newspaper, yend=fit$fitted), color='red', alpha
        =0.5) +
    theme_minimal()
```



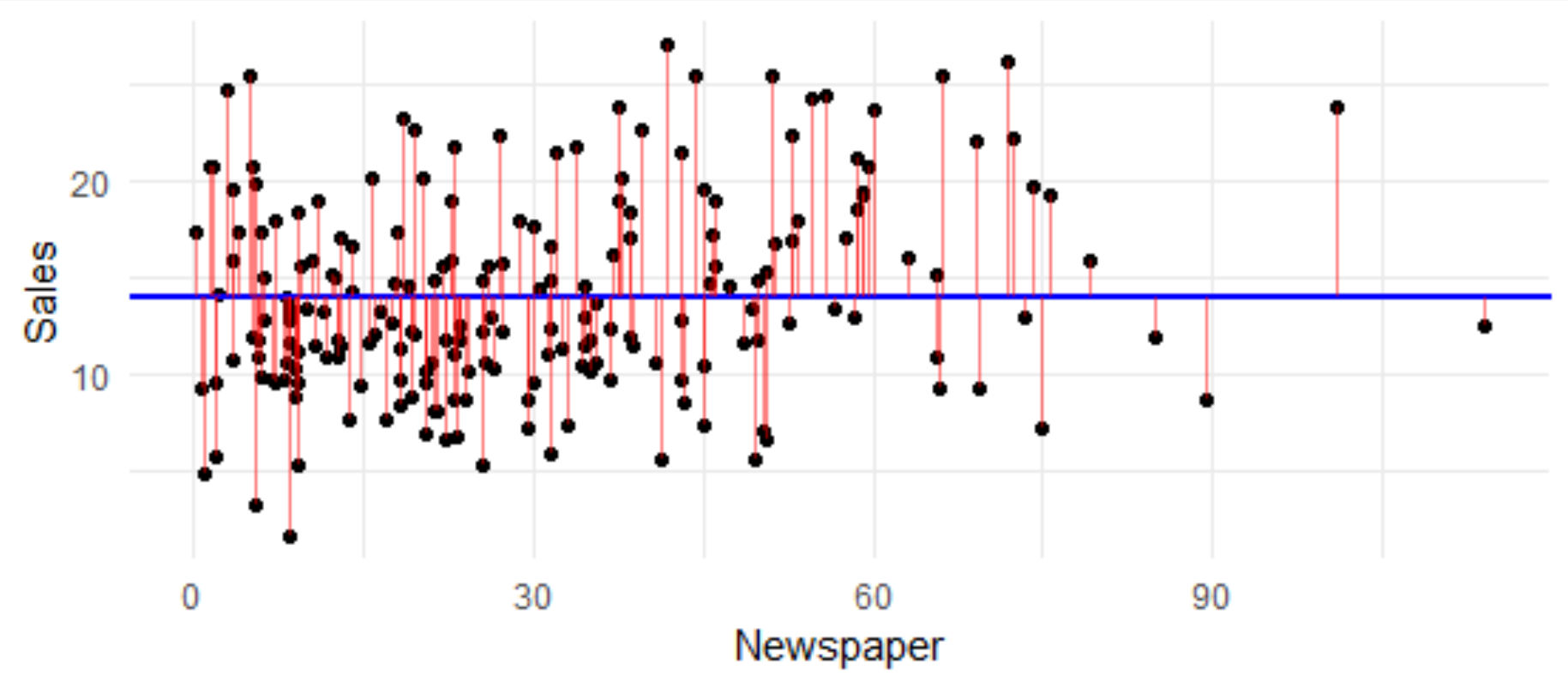
R^2 – Example: Comparison

In the plot above, the blue line has the property that the sum of the squared red lines is minimal. Let's compare the sum of squared errors for this optimal fit to the fit with the line having the mean value.

```
summary(fit)$df[2]  # (n - p - 1)
summary(fit)$sigma * summary(fit)$df[2] # 1008.311112571
m <- mean(adv$Sales)  # 14.0225
sum((adv$Sales - m)^2)  # 5417.149
```

R^2 – Example: Suboptimal fit

```
ggplot(adv, aes(Newspaper, Sales)) +
    theme_minimal() +
    geom_point() +
    geom_abline(slope=0, intercept=m, color='blue', size=1) +
    geom_segment(aes(xend=Newspaper, yend=m), color='red', alpha=0.5)
```



R^2 – Definition

- σ^2 is the error variance, the variability in Y which is not explained by $X\beta$.
- $X\hat{\beta}$ explains most variability in Y.
- A measure for goodness of the fit with the linear model is the *coefficient of determination* R^2 which is defined by

$$R^2 := \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}. \tag{13}$$

- *TSS* is the *total sum of squares* which measures the total variance before the regression (see previous slide), defined by

$$\text{TSS} := \sum_{i=1}^{n} (Y_i - \overline{Y})^2. \tag{14}$$

- *RSS* is the *residual sum of squares* which measures the variability after performing the regression, defined by

$$\text{RSS} := ||\boldsymbol{y} - \boldsymbol{X}\hat{\beta}||^2. \tag{15}$$

R^2 – Interpretation

- By definition, R^2 is the proportion of the total variability minus the variability after the regression, in relation to the total variability.
- R^2 has values between 0 and 1.
- A value close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- A value close to 0 indicates that 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 high, or both.
- In our example, we have

$$R^2 \approx 1 - \frac{1008.311}{5417.149} \approx 0.814,$$

meaning that approximately 81.4% of the variability have been explained by the regression.

Adjusted R^2

- By including more (not perfectly collinear) predictors into the model will always increase explained variation.
- The *adjusted* R^2, denoted as $\overline{R}^2$, measures as R^2 also how much variability have been explained by the regression, but also takes into accounts the number of predictors. It is defined as

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}\,. \tag{16}$$

- $\overline{R}^2$ is smaller than R^2 if $\text{RSS} \neq 0$ and $p > 0$.
- As smaller p make inference easier, one should choose p such that $\overline{R}^2$ is the largest.

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 have the following properties, among others:
  1. The values in the residual plot are scattered around zero without a visible trend $\Rightarrow$ model assumption is reasonable.
  2. The values in the residual plot exhibit a visible trend/pattern $\Rightarrow$ model assumption is NOT reasonable.
  3. The scatter plot or residual plot exhibits unusual values being far away from most of data $\Rightarrow$ Outliers!
  4. The magnitudes of the measurement errors are not roughly constant across observations $\Rightarrow$ Heteroskedasticity (variance heterogeneity).

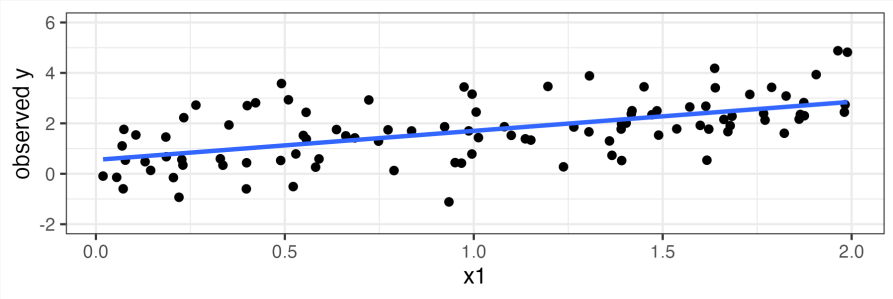
Residual plots: Data set 1

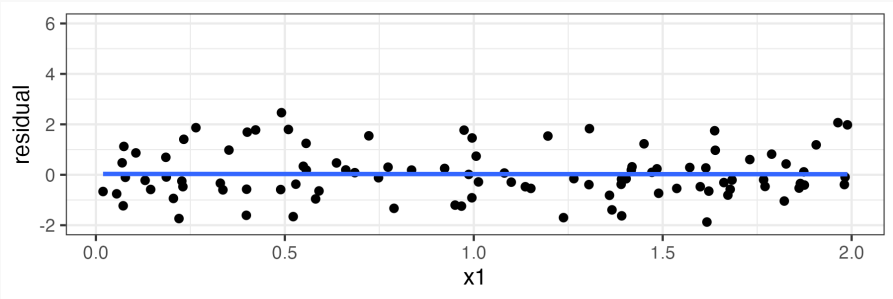
We create the following data consisting of 100 rows.

```
# 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
```

Residual plots: Data set 1

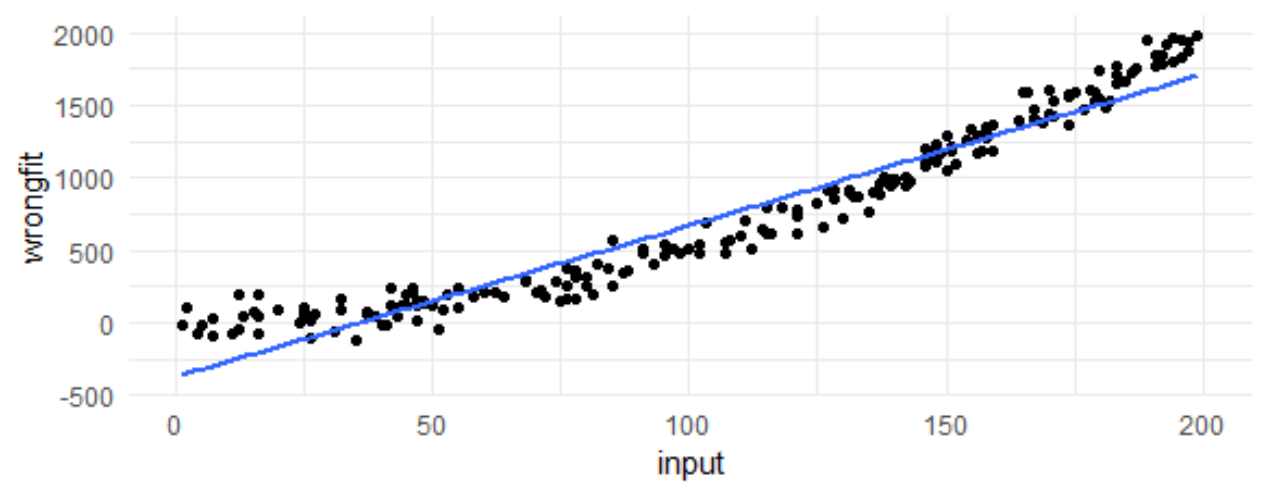
```
library(ggplot2)
# Plot line of best fit
ggplot(df1, aes(x1, yobs)) +
  geom_point() +
  geom_smooth(method='lm', se=F) +
  scale_y_continuous(limits=c(-2,6))+
  labs(y='observed y') +
  theme_minimal()

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

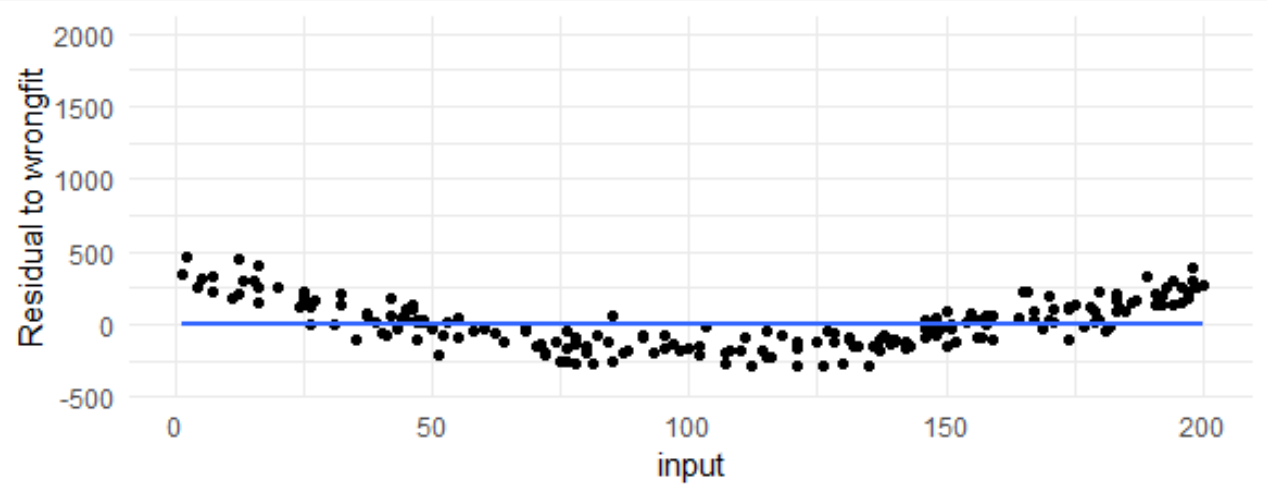




No visible pattern in residual plot $\Rightarrow$ Proper fit of the data using a linear model

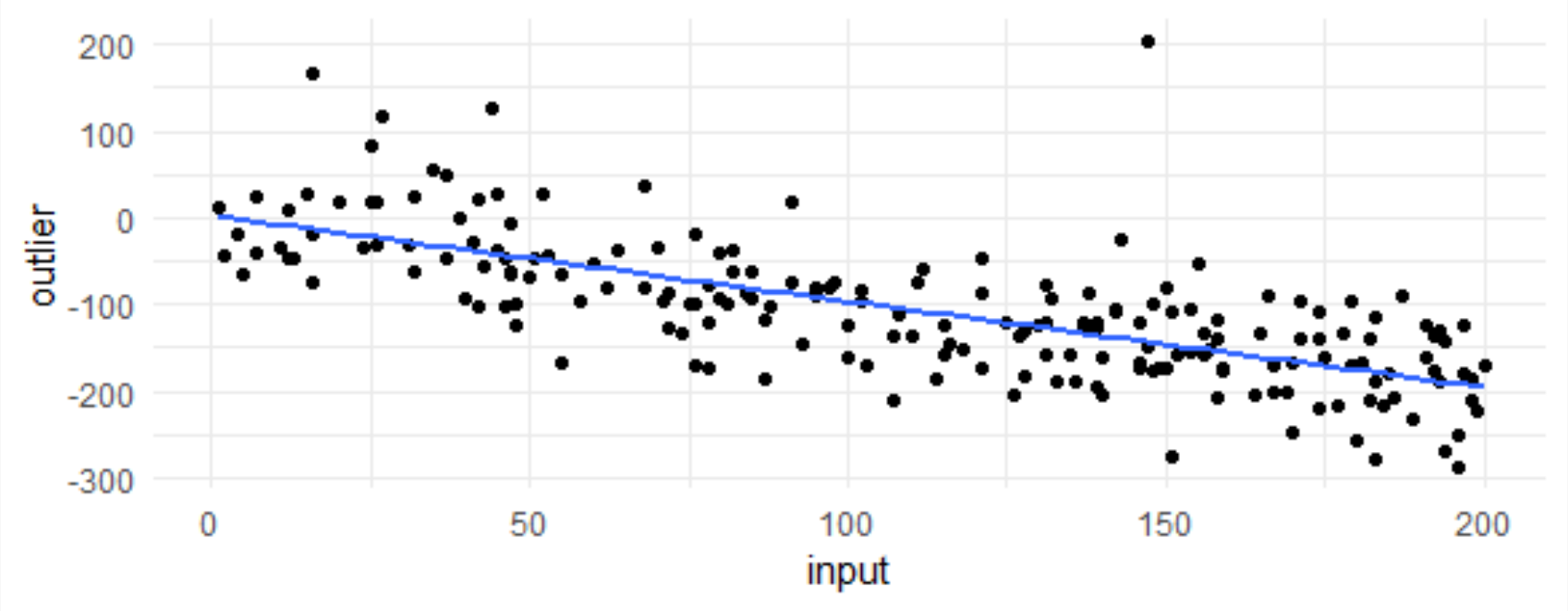
Residual plots: Data set 2





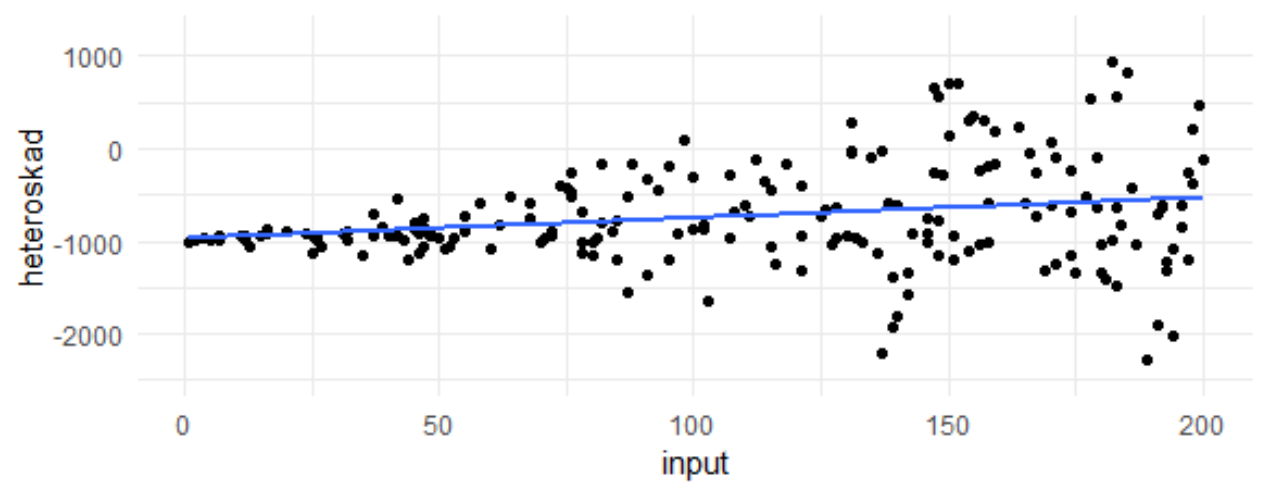
Visible pattern in the residual plot $\Rightarrow$ The data are not properly fitted by the linear model. Maybe a quadratic relationship is reasonable.

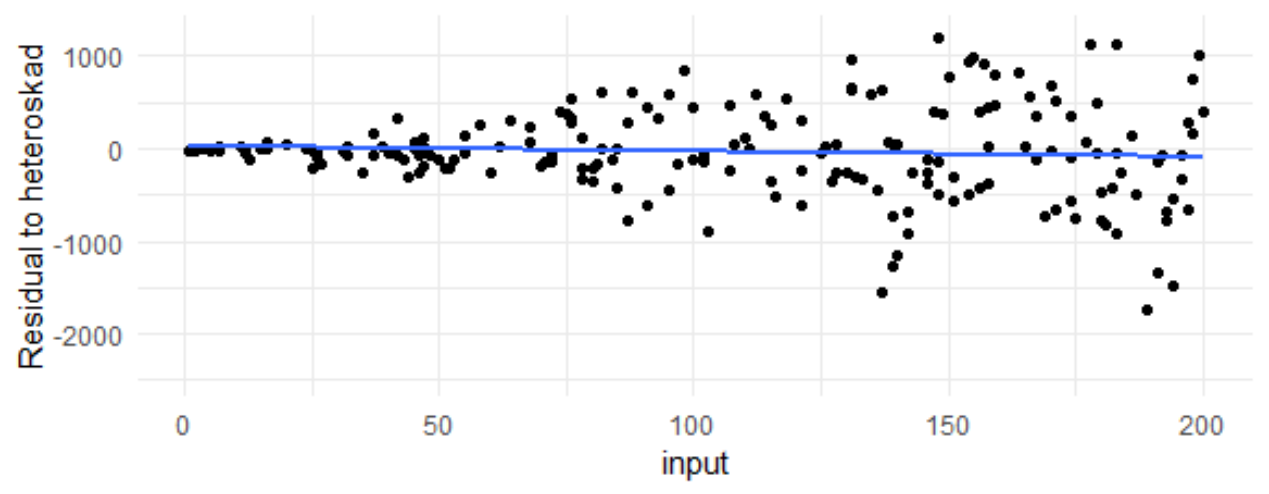
Residual plots: Data set 3



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

Residual plots: Data set 4





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

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 $x_{n+1,1}, \ldots, x_{n+1,p}$. (How?)

Three types of uncertainty associated with this prediction:

1. Inaccuracy in the coefficient estimates $\hat{\beta}$ — 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 incorporates both model uncertainty and observation variance.

Section 6: Regression Analysis with R

Idea of polynomial regression

Idea

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 polynomial function

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \cdots + \beta_d X_1^d + \varepsilon, \tag{17}$$

  where $d \in \mathbb{N}$ is the degree of the poynomial.
- 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

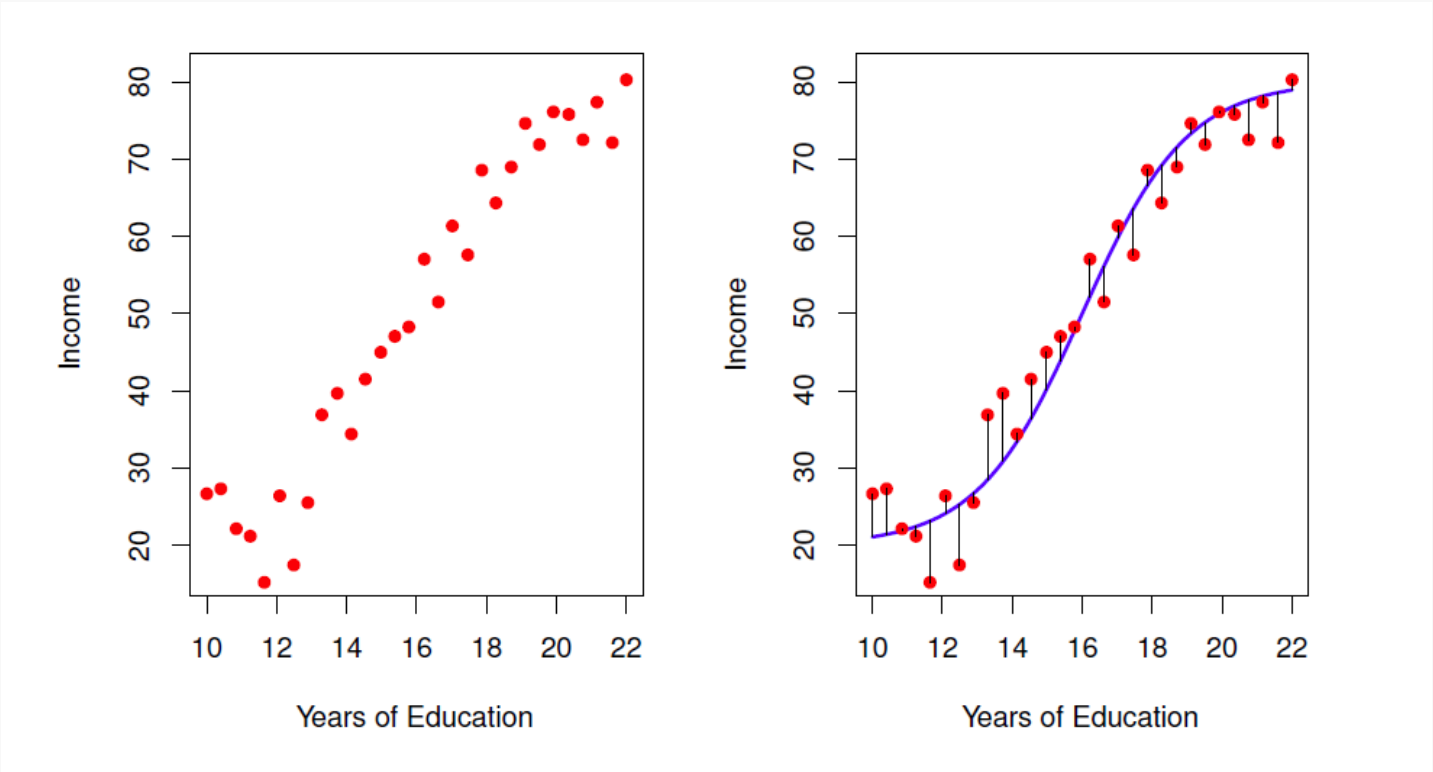


Figure 3: Image by James et al. (2021), based on the `Income` data set in R. The red dots are the observed values of `income` in tens of thousand dollars and `years of education` for 30 individuals.

Example 2: degree-4 polynomial

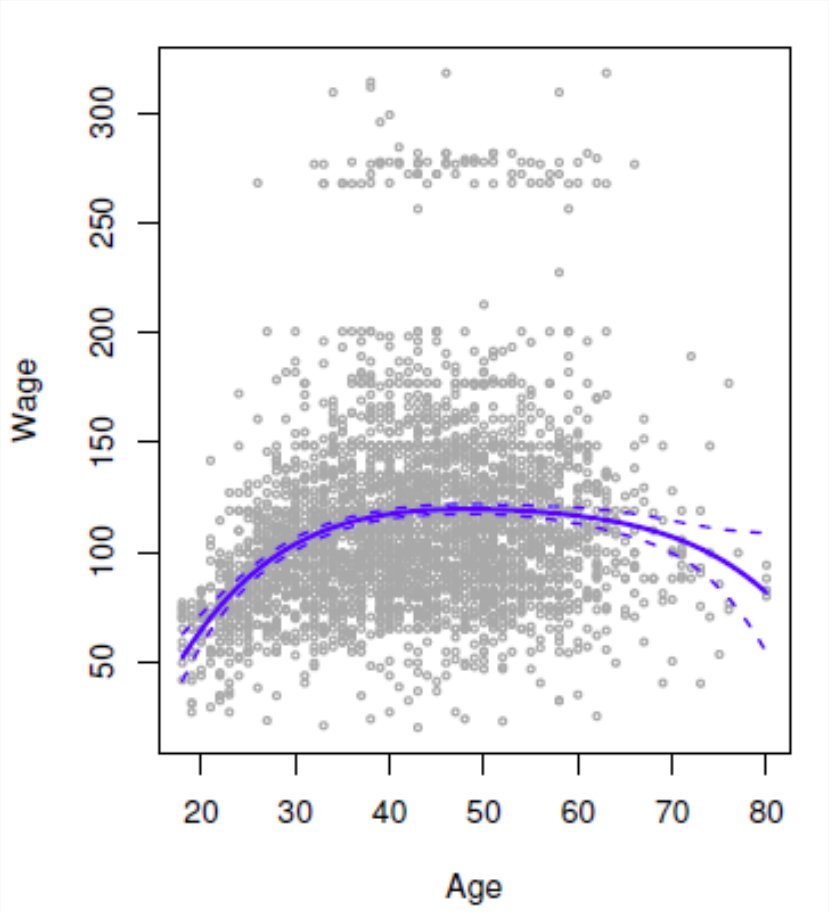


Figure 4: Image by James et al. (2021). The solid blue curve is a degree-4 polynomial of wage (in thousands of dollars) as a function of age, fit by least squares.

Example 3: Polynomial regression with two predictors

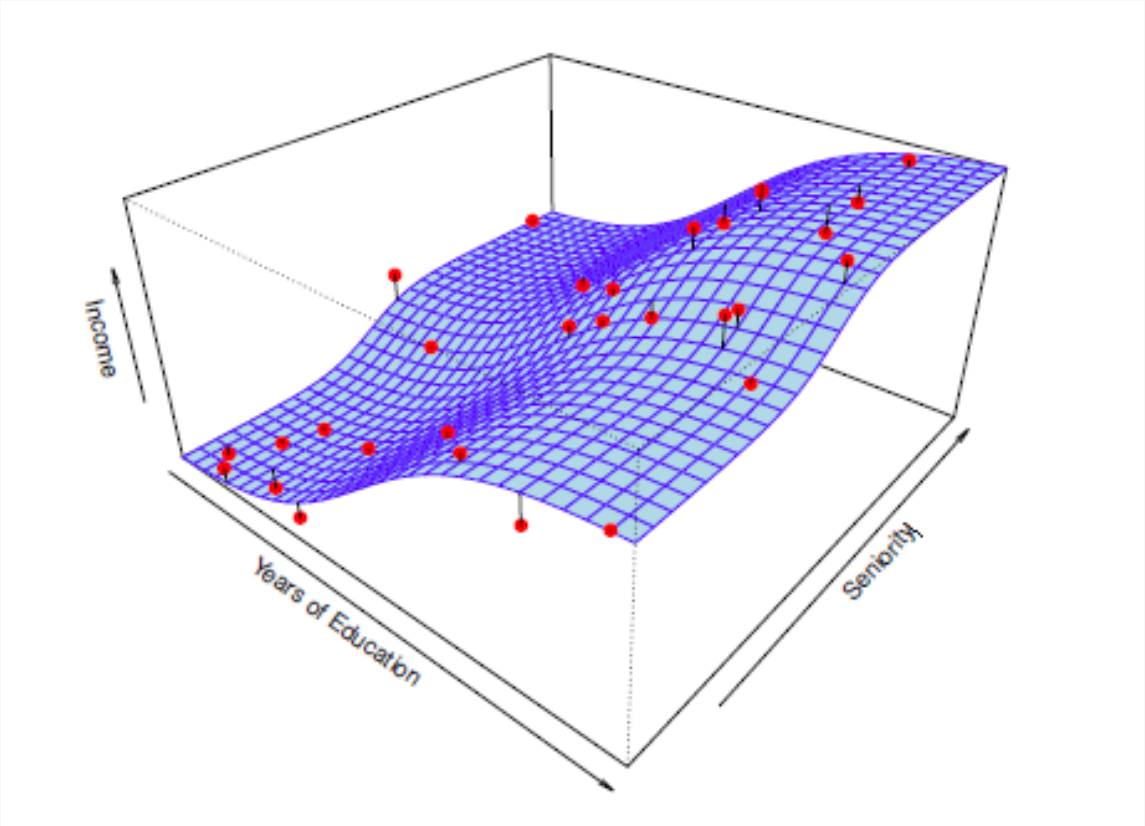


Figure 5: Image by James et al. (2021), based on the `Income` data set in R. The `income` is displayed as a function of `years of education` and `seniority`, where linearity does not seem appropriate. It might be reasonable to do polynomial regression with two predictors.