Linear Regression and its p-values

Statistics modeling group

Gabriel Raya
Tilburg University

Objectives

- **Statistical modeling** : to reason about the set of **statistical assumptions** and limitation underpinning linear regression.
- **Data generative process** : to think about how our statistical model represents the data generative process and to discuss weather this makes and when.
- **How these two influence the way test statistics are constructed**

Linear Regression

- It assumes that the dependence between Y on p predictors $X_1, X_2, \ldots, X_p$ is linear

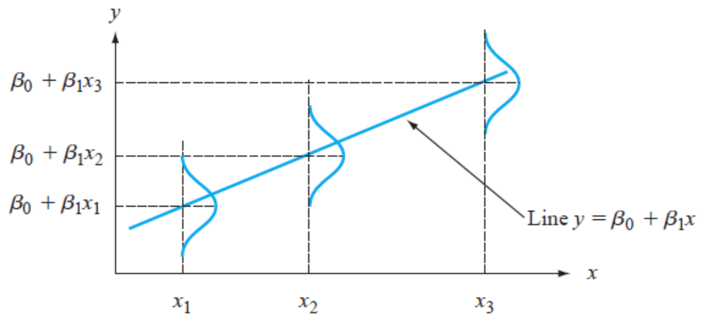
$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i$$

- True regression functions are never linear!
- However, it serves as a good interpretable approximation

Simple Linear Regression

- **Model:** $$Y = \beta_0 + \beta_1 X + \epsilon$$

  ϵ is the random error so Y is a random variable too.

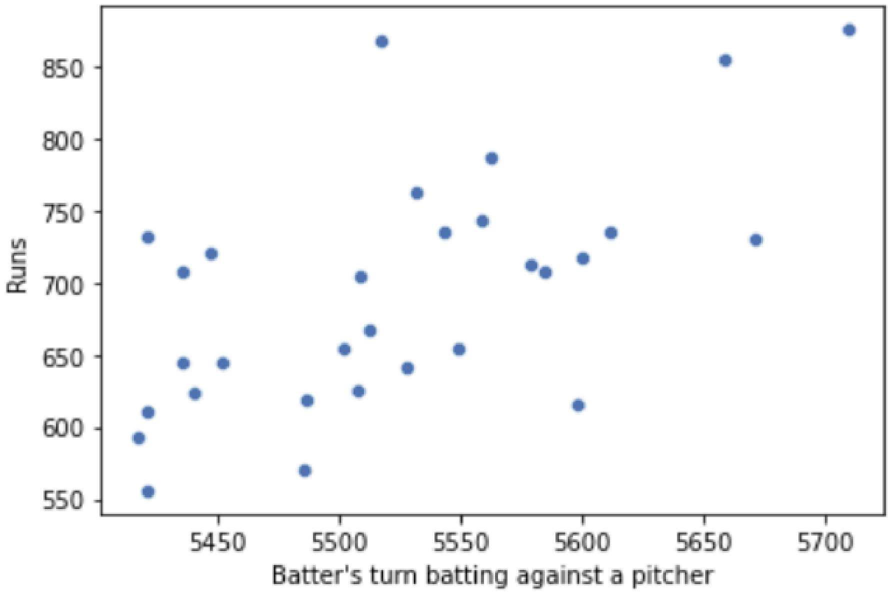


Distribution of Y for different values of x

- **Data:** $$(X_1, Y_1), \dots, (X_n, Y_n) \sim F_{X,Y}$$

  each (X_i, Y_i) satisfies $Y_i = \beta_0 + \beta_1 X_i + \epsilon$

Example

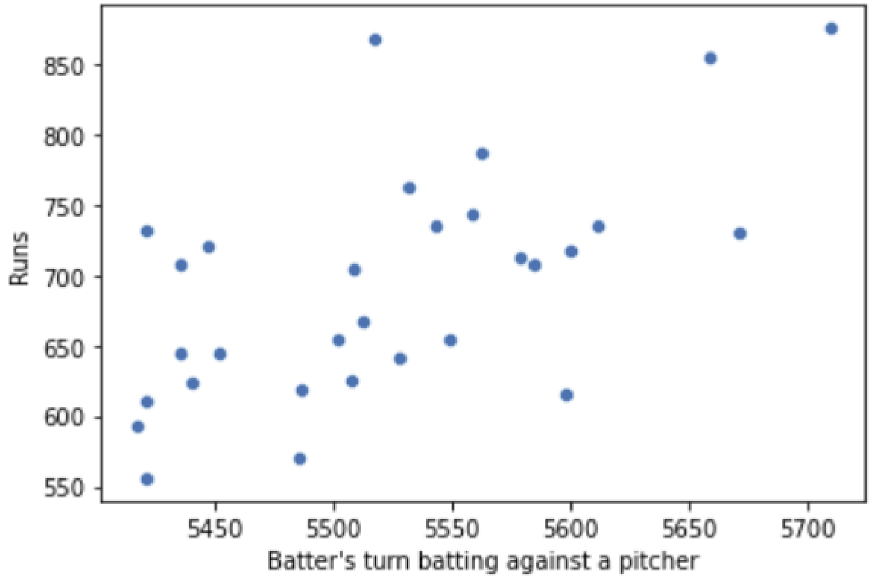


Would you fit a linear model ? Why?

Checking Assumptions

The following assumption should be met to before fitting a linear model

1. **Direction** (positive, negative)
2. **Form** (linear or not linear)
3. **Strength** of the association between the explanatory and response numerical variables



The relationship between the two variables seems **linear, upward sloping, and moderately strong.**

Hypothesis testing on linear regression

- Standard errors are used to perform ***hypothesis testing*** on the coefficients.

H_0: There is no relationship between X and Y

H_A: There is some relationship between X and Y

- Mathematically, for this corresponds to testing

$$H_0:\ \beta_1 = 0$$

versus

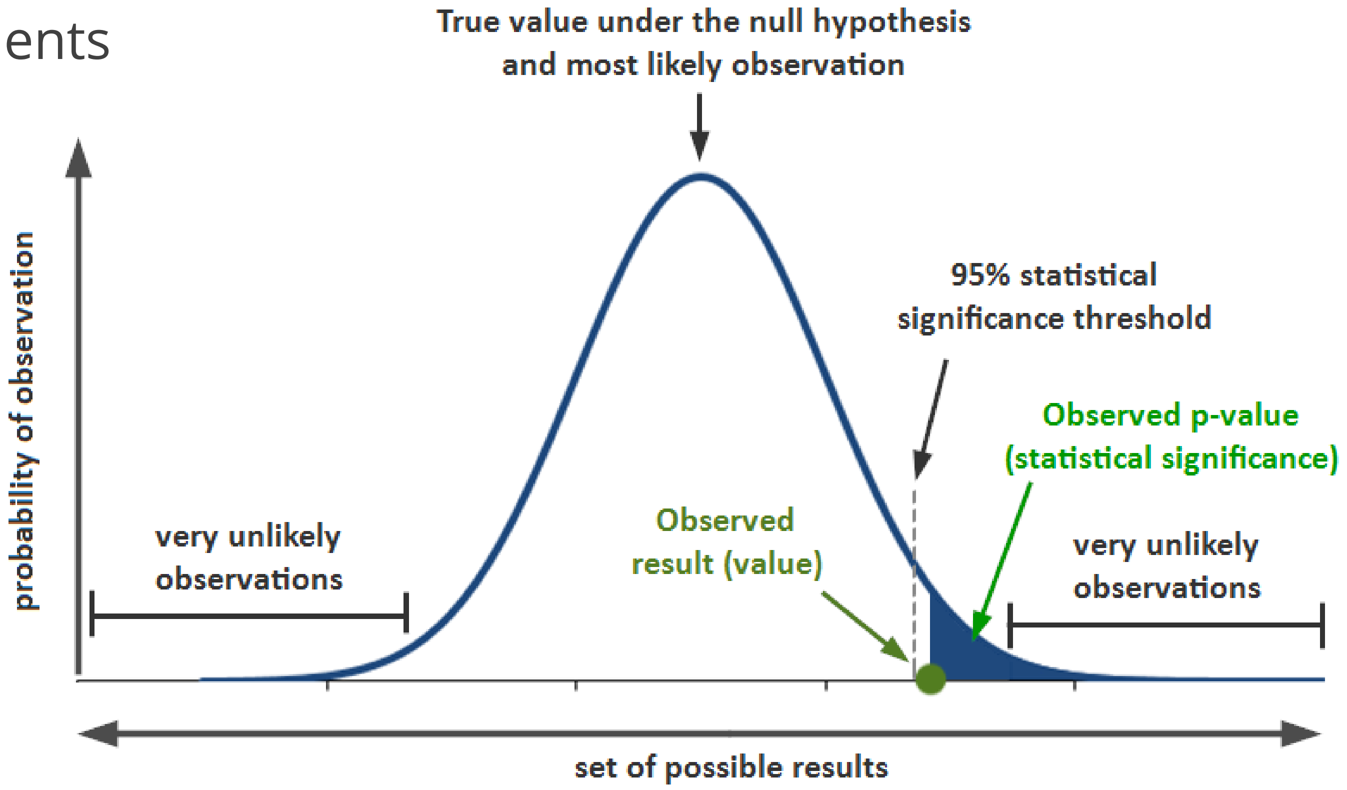
$$H_A:\ \beta_1 \neq 0$$

How do we reject H_0 ?

The decision to reject H_0 is based on the critical value α and the test statistics $T(X)$

Therefore, we need two components

- A **test statistics** $\boldsymbol{T(X)}$
- A **critical value** α

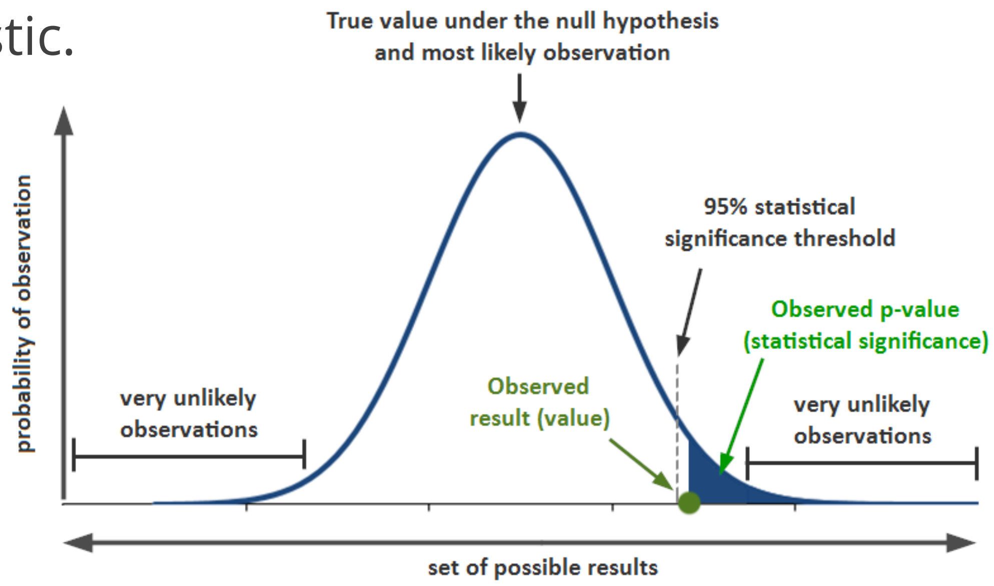


P-value

- The p-value indicates the probability of obtaining a value equal or more extreme than the observed test statistic, given that H_0 is true.

$$P(T(X) \geq \ t_{obs} \,|H_0 = 1)$$

Where t_{obs} is the observed value of the test statistic.



How do we construct this T(X)?

Hypothesis testing

To test the null hypothesis, we compute a **t-statistic**, given by

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- This will have a ***t-distribution*** with n-2 degrees of freedom, assuming $\beta_1 = 0$

Why a t-distribution?

Modeling assumptions and implications

Unknown parameters

- Intercept $\boldsymbol{\beta_0}$
- Slope $\boldsymbol{\beta_1}$
- Variance σ^2

Estimated model $\widehat{r(x)} = \hat{\beta}_0 + \hat{\beta}_1 x$

Predicted values $Y_i = \widehat{r(X_i)}$

Residuals $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} = Y_i - \hat{Y}_i = \boldsymbol{Y_i} - \left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} \boldsymbol{x}\right)$

Residuals Sum of Squares (RSS) $RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}{}^2$

Least Squares Estimator

- The least Squares estimates are $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the $RSS = \sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}{}^2$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\left(\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}{}^2\right)$$

- After optimizing we obtain:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = \frac{\sum_{i=1}^{n}(\boldsymbol{X}_{\boldsymbol{i}} - \overline{\boldsymbol{X}})(\boldsymbol{Y}_{\boldsymbol{i}} - \overline{\boldsymbol{Y}})}{\sum_{i=1}^{n}(\boldsymbol{X}_{\boldsymbol{i}} - \overline{\boldsymbol{X}_{\boldsymbol{i}}})^{2}} = \frac{\boldsymbol{S}_{\boldsymbol{XY}}}{\boldsymbol{S}_{\boldsymbol{XX}}} = \frac{\text{Sample Covariance between X and Y}}{\text{Sample Variance of X}}$$

$$\widehat{\boldsymbol{\beta}}_{\mathbf{0}} = \overline{\boldsymbol{Y}} - \widehat{\boldsymbol{\beta}}_{\mathbf{1}}\overline{\boldsymbol{X}}$$

Estimating the Variance σ^2

The variance σ^2 of the random error ϵ is usually not known. So, it is necessary to estimate it!

- An **unbiased estimate** for $\boldsymbol{\sigma^2}$ is

$$\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}{}^2$$

Proof

- $\frac{RSS = \sum_{i=1}^{n} \hat{\epsilon}_i{}^2}{\sigma^2} \sim \chi^2(n-2)$ distribution with n-2 degrees of freedom [see Cochran's theorem]
- **Property :** if $w \sim \chi^2(n) \;\rightarrow E[w] = n.$

$$E[\hat{\sigma}^2] = E\left[\left(\frac{1}{n-2}\right)\sum_{i=1}^{n} \hat{\epsilon}_i{}^2\right] = \left(\frac{1}{n-2}\right) E\left[\sum_{i=1}^{n} \hat{\epsilon}_i{}^2\right] = \left(\frac{1}{n-2}\right)\sigma^2(n-2) = \sigma^2$$

Variance of estimators $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_0$

Since $\boldsymbol{\sigma}^2$ is unknown, we use $\hat{\sigma}^2$

$$\mathrm{V}[\hat{\beta}_0] = \frac{\hat{\sigma}^2 \cdot \sum_{i=1}^{n} {X_i}^2}{n \cdot S_{XX}}$$

$$V[\hat{\beta}_1] = \frac{\hat{\sigma}^2}{S_{XX}}$$

Note: $\mathrm{V}[\hat{\beta}_0]$ and $\mathrm{V}[\hat{\beta}_1]$ are also known as the standard error $SE^2(\hat{\beta}_0)$ and $SE^2(\hat{\beta}_1)$.

Implications of properties of random error for the Estimators

- **Assumptions on the Random Error ϵ**
  - $\mathrm{E}[\epsilon_i] = 0$; $\mathrm{V}[\epsilon_i] = \sigma^2$; **Each ϵ_i is i.i.d. normally distributed.**

- **Implications for the Response Variable Y**
  - $E[Y_i] = E[\beta_0 + \ \beta_1 X_i + \epsilon\ ] = \beta_0 + \beta_1 X_i$
  - $\mathrm{V}[Y_i] = \sigma^2$
- **Implications for the Estimators**
  - $\hat{\beta}_1$ is normally distributed with mean β_1 and variance $\frac{\hat{\sigma}^2}{S_{XX}}$
  - $\hat{\beta}_0$ is normally distributed with mean β_0 and variance $\frac{\hat{\sigma}^2 \cdot \sum_{i=1}^{n} {X_i}^2}{n \cdot S_{XX}}$
  - $\frac{RSS}{\sigma^2} \sim \chi^2(n-2)$ distribution with n-2 degrees of freedom
  - $\hat{\sigma}^2 \sim \chi^2(n-2)$ distribution with n−2 degrees of freedom
  - $\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$

$\boldsymbol{\hat{\beta}_k \sim N(0,1)}$ **and** $\boldsymbol{\hat{\sigma}^2 \sim \chi^2(n-2)}$

Testing of hypotheses for slope parameter

Case : When σ^2 is unknown:

- We know that
  - $\frac{RSS}{\sigma^2} \sim \chi^2(n-2)$
  - $E\left[\hat{\sigma}^2 = \left(\frac{1}{n-2}\right)\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\,2}\right] = \sigma^2$
  - $\frac{RSS}{\sigma^2}$ and $\hat{\beta}_1$ are independently distributed
- Thus, the following statistic can be constructed:
- $t_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}}$

T-distribution

[Student's t distribution] Let $Z \sim N(0,1)$ and $W \sim \chi^2(n)$ be independent random variables. The random variable

$$T = \frac{Z}{\sqrt{W/n}}$$

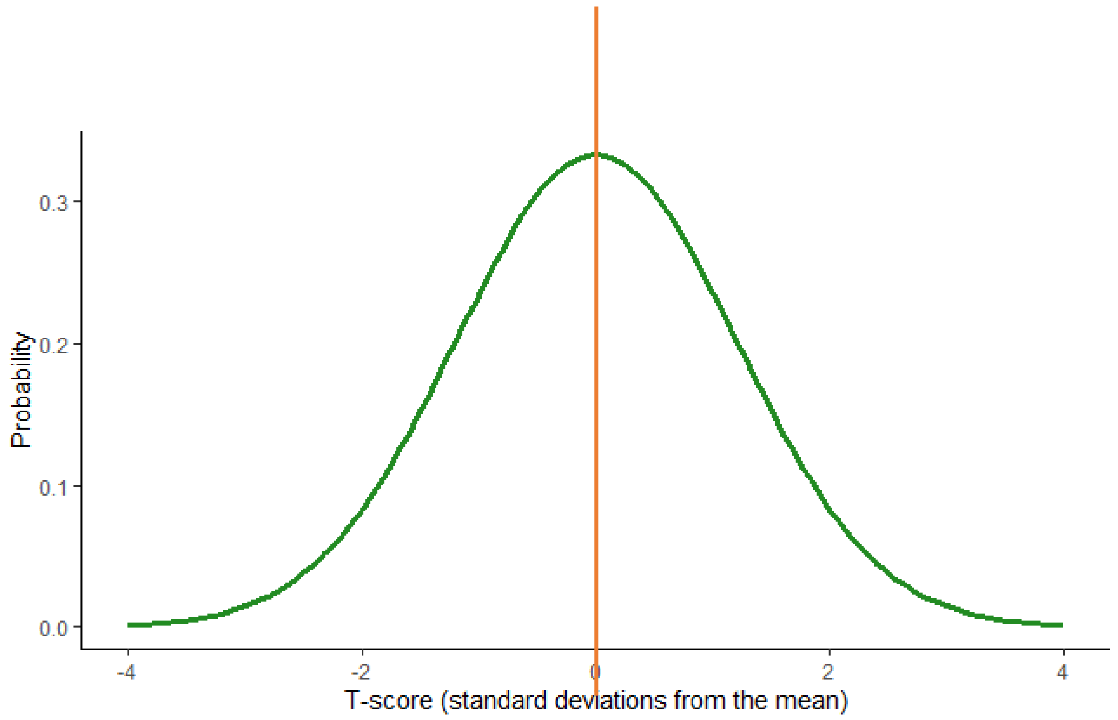
has by definition a (Student's) t distribution with n degrees of freedom, denoted by

$$T \sim t(n).$$

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}}$$

Interpretation of the t-score

- For a random variable this will typically refer to the ratio of the departure of the estimated value of a parameter from its hypothesized value to its standard error.
- Tell us how many standard deviations our observed value is from the mean (null hypothesis)



$\hat{\beta}_k$=0

When should we use a t-test

- Small sample size or
- Unknown population standard deviation

Confidence Intervals

- The standard error on an estimator reflects how it varies under repeating sampling.
  - It can be used to construct a **Confidence Interval**

$$\hat{\beta}_k \pm t \; * SE(\hat{\beta}_k)$$

Assessing the Overall Accuracy of the Model R-square

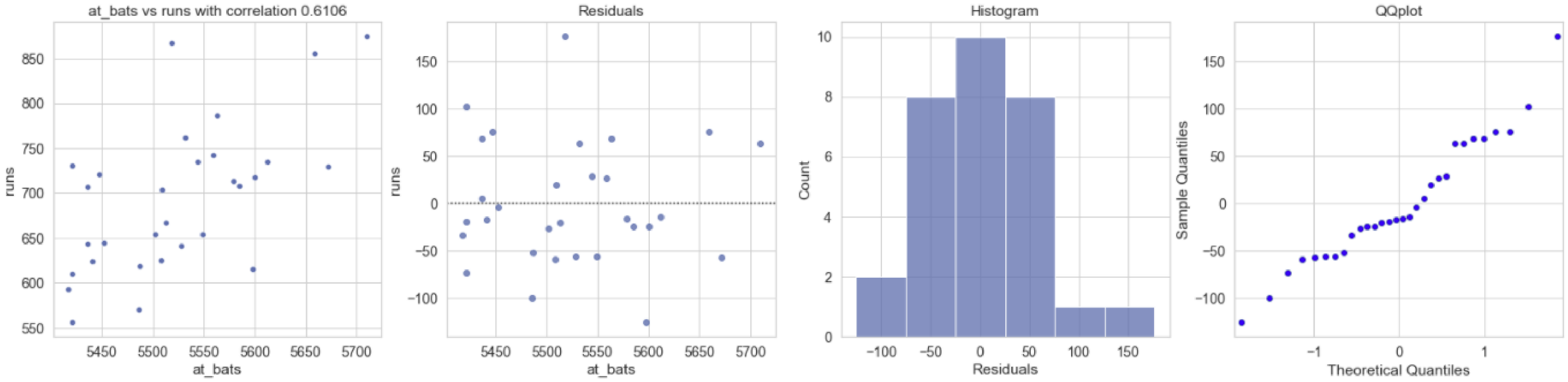
- R-squared or fraction of variance explained is

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

- Where $TSS = \sum_{i=1}^{n}(Y_i \ - \ \bar{Y})$

Model Diagnosis

1. **Linearity,**
2. **Nearly normal residuals, and**
3. **Constant variability.**



Multiple linear regression

- Adjusted R^2 instead of R^2

- **Model selection**
  - Forward method
  - Backward method

Backward selection with the p-value

1. Starts with the **full model.**
2. Drop the variable with the highest **p-value** and refit a smaller model
3. Repeat until all variables left in the model are statistically significant

Results

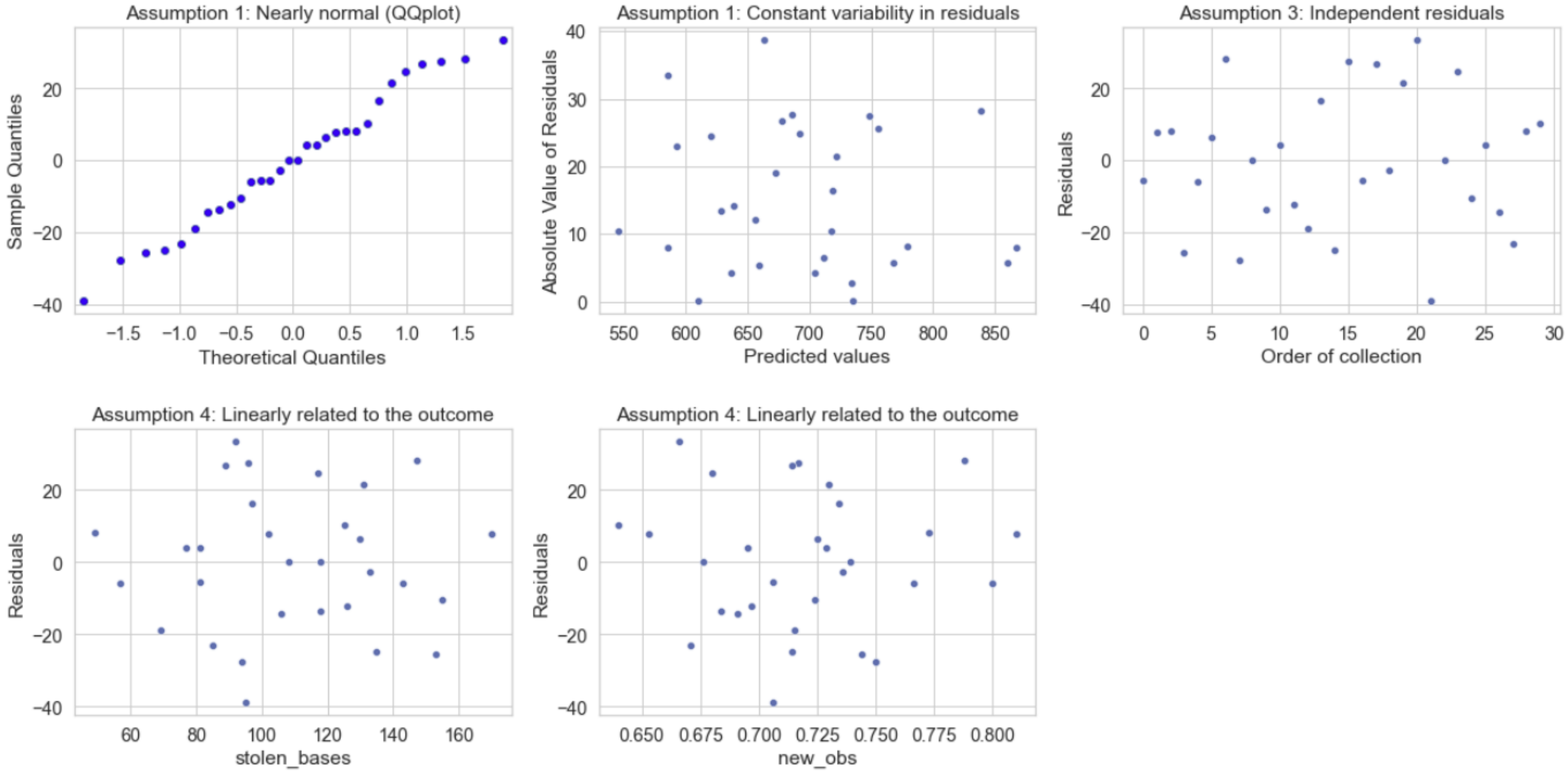
- After running the backward selection method, we obtained

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-731.1710	65.095	-11.232	0.000	-864.735	-597.607
stolen_bases	0.3154	0.122	2.596	0.015	0.066	0.565
new_obs	1933.3858	87.346	22.135	0.000	1754.167	2112.604

Model diagnostics

1. **Residuals** of the model are **nearly normal**
2. **Variability** of the **residuals is nearly constant**
3. **Residuals** are **independent**
4. **Each variable is linearly dependent to the outcome**

Results


Assumption 1: Nearly normal (QQplot)
Sample Quantiles
Theoretical Quantiles
Assumption 1: Constant variability in residuals
Absolute Value of Residuals
Predicted values
Assumption 3: Independent residuals
Residuals
Order of collection
Assumption 4: Linearly related to the outcome
Residuals
stolen_bases
Assumption 4: Linearly related to the outcome
Residuals
new_obs

References

- Devore, J. L. (2011). *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole.
- Larry Wasserman. 2010. All of Statistics: A Concise Course in Statistical Inference. Springer Publishing Company, Incorporated.
- Slides used from other people:
  - http://home.iitk.ac.in/~shalab/econometrics/Chapter2-Econometrics-SimpleLinearRegressionAnalysis.pdf
  - https://web.stanford.edu/~hastie/MOOC-Slides/linear_regression.pdf
  - http://www.unm.edu/~lspear/geog525/24linreg2.pdf
  - dr. L. Geerligs, Frequentist Statistics slides.
  - Prof. EA Cator (Eric), Regression Analysis and non-parametric statistics

Proofs

$$\frac{dRSS(\hat\beta_0,\hat\beta_1)}{d\hat\beta_0} = \frac{dRSS\left(\sum_{i=1}^{n}\left(Y_i - (\hat\beta_0 + \hat\beta_1 X_i)\right)^2\right)}{d\hat\beta_0} = \sum_{i=1}^{n} 2\left(Y_i - (\hat\beta_0 + \hat\beta_1 X_i)\right)(-1) = 0$$

$$= \sum_{i=1}^{n}\left(Y_i - (\hat\beta_0 + \hat\beta_1 X_i)\right) = \boldsymbol{\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat\beta_0 - \sum_{i=1}^{n} \hat\beta_1 X_i = 0}$$

$$\frac{dRSS(\hat\beta_0,\hat\beta_1)}{d\hat\beta_1} = \frac{dRSS\left(\sum_{i=1}^{n}\left(Y_i - (\hat\beta_0 + \hat\beta_1 X_i)\right)^2\right)}{d\hat\beta_1} = \sum_{i=1}^{n} 2\left(Y_i - (\hat\beta_0 + \hat\beta_1 X_i)\right)(-X_i) = 0$$

$$\sum_{i=1}^{n}\left(Y_i - (\hat\beta_0 + \hat\beta_1 X_i)\right)(X_i) = \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} \hat\beta_0 X_i - \sum_{i=1}^{n} \hat\beta_1 {X_i}^2 = 0$$

Rearranging we obtain what is called the **normal equations**

$$\sum_{i=1}^{n} Y_i = n\hat\beta_0 + \hat\beta_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = \hat\beta_0 \sum_{i=1}^{n} X_i + \hat\beta_1 \sum_{i=1}^{n} {X_i}^2$$

Proofs

Normal Equations $\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i & (1) \\ \sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} {X_i}^2 & (2) \end{cases}$

By solving eq (1) for $\hat{\beta}_0$ we obtained

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

By substituting $\hat{\beta}_0$ on Equation we get

$$\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} Y_i}{n} \sum_{i=1}^{n} X_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} {X_i}^2 - \frac{\sum_{i=1}^{n} X_i}{n} \sum_{i=1}^{n} X_i \right)$$

By realizing that the left size is S_{XY} and the right-hand side is S_{XX}

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_i)(Y_i - \bar{Y}_i)}{\sum_{i=1}^{n} (X_i - \bar{X}_i)^2} = \frac{S_{XY}}{S_{XX}}$$

Unbiased property

From $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X_i})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \overline{X_i})^2}$ we express $\hat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i$ where $k_i = \frac{\sum_{i=1}^{n}(X_i - \overline{X_i})}{\sum_{i=1}^{n}(X_i - \overline{X_i})^2}$

- Notice that
  - $\bar{Y} \sum_{i=1}^{n}(X_i - \overline{X_i}) = \bar{Y} \cdot 0 = 0$
  - $\sum_{i=1}^{n} k_i = 0$
  - $\sum_{i=1}^{n} k_i X_i = 1$
- $E\left[\hat{\beta}_1\right] = E[\sum_{i=1}^{n} k_i Y_i] = \sum_{i=1}^{n} k_i E[Y_i] = \sum_{i=1}^{n} k_i E[\beta_0 + \beta_1 X_i + \epsilon] = \sum_{i=1}^{n} k_i(\beta_0 + \beta_1 X_i) = \beta_1$