Lecture 4: Linear Regression and Classification

Reading: Chapter 3 and Chapter 4

STATS 202: Data mining and analysis

Rajan Patel

Potential issues in linear regression

1. Interactions between predictors
2. Non-linear relationships
3. Correlation of error terms
4. Non-constant variance of error (heteroskedasticity).
5. Outliers
6. High leverage points
7. Collinearity

Correlation of error terms

We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma) \text{ i.i.d.}$$

What if this breaks down?

Correlation of error terms

We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma) \text{ i.i.d.}$$

What if this breaks down?

The main effect is that this invalidates any assertions about Standard Errors, confidence intervals, and hypothesis tests:

Example: Suppose that by accident, we double the data (we use each sample twice).

Correlation of error terms

We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \varepsilon_i \quad ; \quad \varepsilon_i \sim \mathcal{N}(0, \sigma) \text{ i.i.d.}$$

What if this breaks down?

The main effect is that this invalidates any assertions about Standard Errors, confidence intervals, and hypothesis tests:

Example: Suppose that by accident, we double the data (we use each sample twice). Then, the standard errors would be artificially smaller by a factor of $\sqrt{2}$.

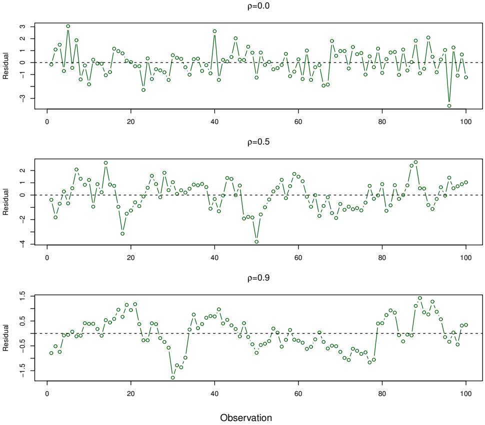
Correlation of error terms

When could this happen in real life:

- **Time series:** Each sample corresponds to a different point in time. The errors for samples that are close in time are correlated.
- **Spatial data:** Each sample corresponds to a different location in space.
- **Predicting height from weight at birth:** Suppose some of the subjects in the study are in the same family, their shared environment could make them deviate from $f(x)$ in similar ways.

Correlation of error terms

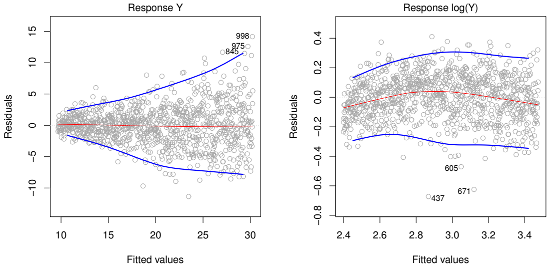
Simulations of time series with increasing correlations between ε_i.



Non-constant variance of error (heteroskedasticity)

The variance of the error depends on the input.

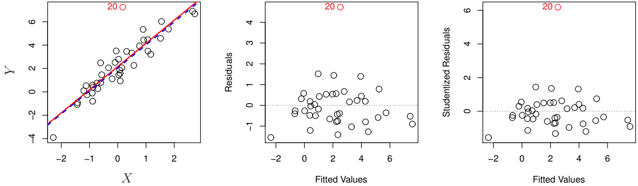
To diagnose this, we can plot residuals vs. fitted values:



Solution: If the trend in variance is relatively simple, we can transform the response using a logarithm, for example.

Outliers

Outliers are points with very high errors.



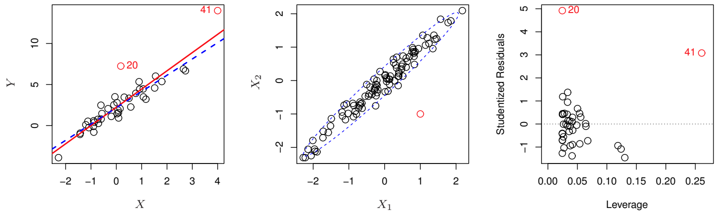
While they may not affect the fit, they might affect our assessment of model quality.

Possible solutions:

- If we believe an outlier is due to an error in data collection, we can remove it.
- An outlier might be evidence of a missing predictor, or the need to specify a more complex model.

High leverage points

Some samples with extreme inputs have an outsized effect on $\hat{\beta}$.

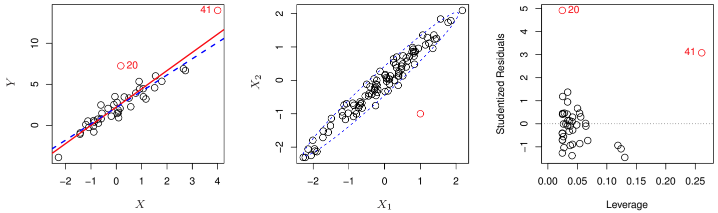


This can be measured with the **leverage statistic** or **self influence**:

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} = (\underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{\text{Hat matrix}})_{i,i} \in [1/n, 1].$$

Studentized residuals

- The residual $\hat{\epsilon}_i = y_i - \hat{y}_i$ is an estimate for the noise ϵ_i.
- The standard error of $\hat{\epsilon}_i$ is $\sigma\sqrt{1 - h_{ii}}$.
- A **studentized residual** is $\hat{\epsilon}_i$ divided by its standard error.
- It follows a Student-t distribution with $n - p - 2$ degrees of freedom.

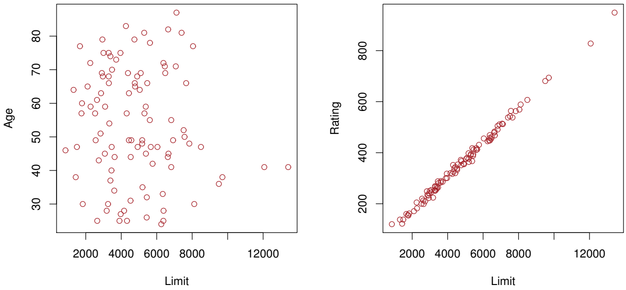


Collinearity

Two predictors are collinear if one explains the other well:

$$\texttt{limit} = a \times \texttt{rating} + b$$

i.e. they contain the same information



Collinearity

Problem: The coefficients become *unidentifiable*.

Collinearity

Problem: The coefficients become *unidentifiable*. Consider the extreme case of using two identical predictors `limit`:

$$\texttt{balance} = \beta_0 + \beta_1 \times \texttt{limit} + \beta_2 \times \texttt{limit}$$

Collinearity

Problem: The coefficients become *unidentifiable*. Consider the extreme case of using two identical predictors `limit`:

$$\begin{aligned}\texttt{balance} &= \beta_0 + \beta_1 \times \texttt{limit} + \beta_2 \times \texttt{limit} \\ &= \beta_0 + (\beta_1 + 100) \times \texttt{limit} + (\beta_2 - 100) \times \texttt{limit}\end{aligned}$$

The fit $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is just as good as $(\hat{\beta}_0, \hat{\beta}_1 + 100, \hat{\beta}_2 - 100)$.

Collinearity

If 2 variables are collinear, we can easily diagnose this using their correlation.

A group of q variables is **multilinear** if these variables "contain less information" than q independent variables. Pairwise correlations may not reveal multilinear variables.

The Variance Inflation Factor (VIF) measures how *necessary* a variable is, or how predictable it is given the other variables:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 statistic for Multiple Linear regression of the predictor X_j onto the remaining predictors.

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.
KNN regression: prototypical nonparametric method.

Comparing Linear Regression to K-nearest neighbors

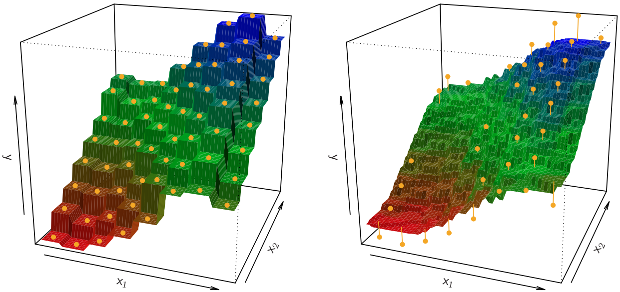
Linear regression: prototypical parametric method.
KNN regression: prototypical nonparametric method.

$$\hat{f}(x) = \frac{1}{K} \sum_{i \in N_K(x)} y_i$$

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.
KNN regression: prototypical nonparametric method.

$$\hat{f}(x) = \frac{1}{K} \sum_{i \in N_K(x)} y_i$$



$K = 1$ $K = 9$

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.

KNN regression: prototypical nonparametric method.

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.

KNN regression: prototypical nonparametric method.

Long story short:

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.

KNN regression: prototypical nonparametric method.

Long story short:

- KNN is only better when the function f is not linear.

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.

KNN regression: prototypical nonparametric method.

Long story short:

- KNN is only better when the function f is not linear.
- When n is not much larger than p, even if f is nonlinear, Linear Regression can outperform KNN.

Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method.

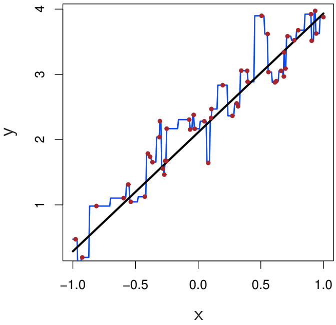
KNN regression: prototypical nonparametric method.

Long story short:

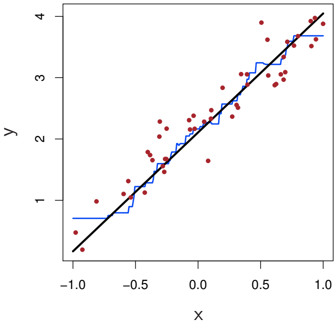
- KNN is only better when the function f is not linear.
- When n is not much larger than p, even if f is nonlinear, Linear Regression can outperform KNN. KNN has smaller bias, but this comes at a price of higher variance.

KNN estimates for a simulation from a linear model

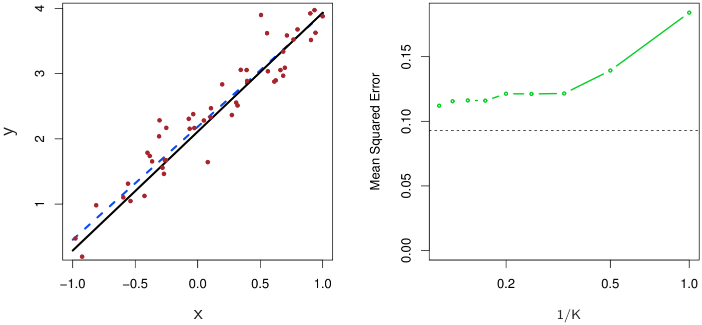
$K = 1$



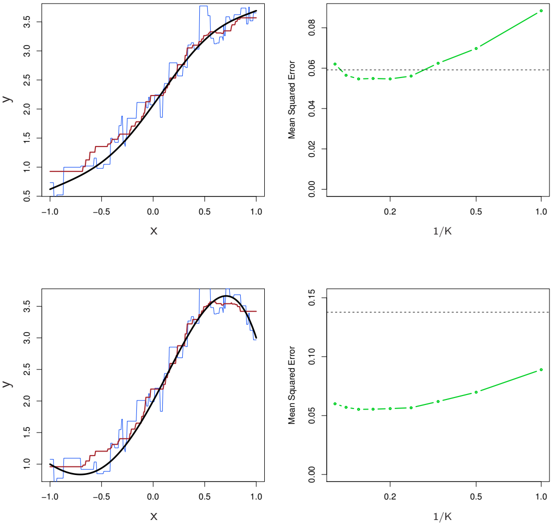
$K = 9$



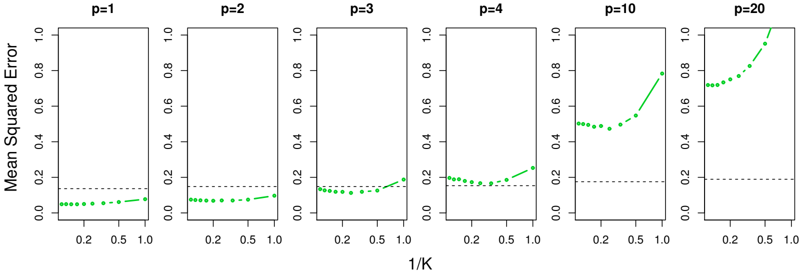
Linear models dominate KNN



Increasing deviations from linearity

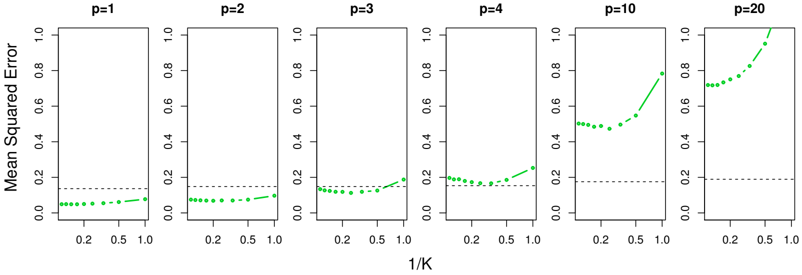


When there are more predictors than observations, Linear Regression dominates



When $p \gg n$, each sample has no nearest neighbors, this is known as the *curse of dimensionality*.

When there are more predictors than observations, Linear Regression dominates



When $p \gg n$, each sample has no nearest neighbors, this is known as the *curse of dimensionality*. The variance of KNN regression is very large.

Classification problems

Supervised learning with a **qualitative or categorical** response.

Just as common, if not more common than regression:

- *Medical diagnosis:* Given the symptoms a patient shows, predict which of 3 conditions they are attributed to.
- *Online banking:* Determine whether a transaction is fraudulent or not, on the basis of the IP address, client's history, etc.
- *Web searching:* Based on a user's history, location, and the string of a web search, predict which link a person is likely to click.
- *Online advertising:* Predict whether a user will click on an ad or not.

Strategy: estimate $P(Y \mid X)$

If we have a good estimate for the conditional probability $\hat{P}(Y \mid X)$, we can use the classifier:

$$\hat{y}_0 = \text{argmax}_{\,y}\, \hat{P}(Y = y \mid X = x_0).$$

Suppose Y is a binary variable. Could we use a linear model?

$$P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_1 X_p$$

Problems:

- This would allow probabilities <0 and >1.
- Difficult to extend to more than 2 categories.

Logistic regression

We model the joint probability as:

$$P(Y = 1 \mid X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}},$$

$$P(Y = 0 \mid X) = \frac{1}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}.$$

This is the same as using a linear model for the log odds:

$$\log \left[\frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)} \right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Fitting logistic regression

The training data is a list of pairs $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$. In the linear model

$$\log\left[\frac{P(Y = 1 \mid X)}{P(Y = 0 \mid X)}\right] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

we don't observe the left hand side.

We cannot use a least squares fit.

Fitting logistic regression

Solution:

The likelihood is the probability of the training data, for a fixed set of coefficients $\beta_0, \dots, \beta_p$:

$$\prod_{i=1}^{n} P(Y = y_i \mid X = x_i)$$

$$= \underbrace{\prod_{i; y_i = 1} \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}}_{\text{Probability of responses} = 1} \underbrace{\prod_{j; y_j = 0} \frac{1}{1 + e^{\beta_0 + \beta_1 x_{j1} + \dots + \beta_p x_{jp}}}}_{\text{Probability of responses} = 0}$$

- Choose estimates $\hat{\beta}_0, \dots, \hat{\beta}_p$ which maximize the likelihood.
- Solved with numerical methods (e.g. Newton's algorithm).

Logistic regression in R

```
> glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,
    data=Smarket,family=binomial)
> summary(glm.fit)

Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5
    + Volume, family = binomial, data = Smarket)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -1.45   -1.20    1.07    1.15    1.33

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.12600    0.24074   -0.52     0.60
Lag1        -0.07307    0.05017   -1.46     0.15
Lag2        -0.04230    0.05009   -0.84     0.40
Lag3         0.01109    0.04994    0.22     0.82
Lag4         0.00936    0.04997    0.19     0.85
Lag5         0.01031    0.04951    0.21     0.83
Volume       0.13544    0.15836    0.86     0.39
```

Logistic regression in R

- We can estimate the Standard Error of each coefficient.
- The z-statistic is the equivalent of the t-statistic in linear regression:

$$z = \frac{\hat{\beta}_j}{\mathsf{SE}(\hat{\beta}_j)}.$$

- The p-values are test of the null hypothesis $\beta_j = 0$.

Example: Predicting credit card `default`

Predictors:

- `student`: 1 if student, 0 otherwise.
- `balance`: credit card balance.
- `income`: person's income.

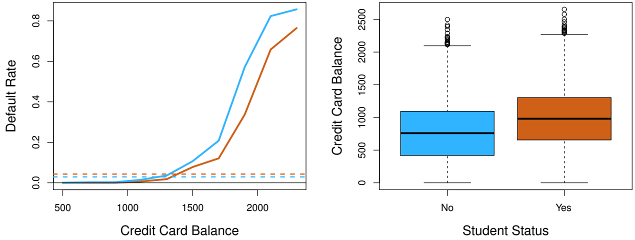
In this dataset, there is *confounding*, but little collinearity.

- Students tend to have higher balances. So, `balance` is explained by `student`, but not very well.
- People with a high `balance` are more likely to default.
- Among people with a given `balance`, students are less likely to default.

Example: Predicting credit card `default`

Predictors:

- `student`: 1 if student, 0 otherwise.
- `balance`: credit card balance.
- `income`: person's income.



Example: Predicting credit card `default`

Logistic regression using only `balance`:

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

Logistic regression using only `student`:

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−3.5041	0.0707	−49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Logistic regression using all 3 predictors:

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

Some issues with logistic regression

- The coefficients become unstable when there is collinearity. Furthermore, this affects the convergence of the fitting algorithm.
- When the classes are well separated, the coefficients become unstable. This is always the case when $p \geq n - 1$.