Lecture 7: Linear Regression (continued)

Reading: Chapter 3

STATS 202: Data mining and analysis

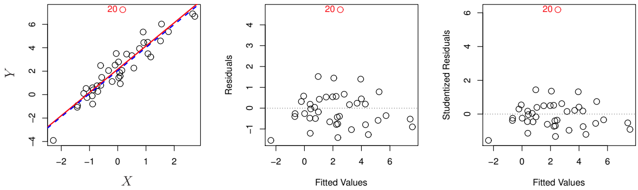
Jonathan Taylor, 10/8
Slide credits: Sergio Bacallado

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

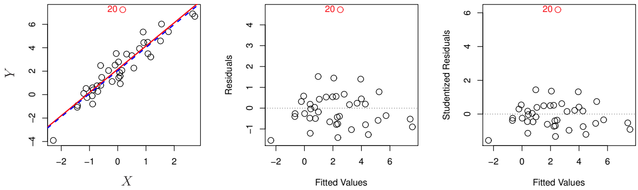
Outliers

Outliers are points with very high errors.



Outliers

Outliers are points with very high errors.

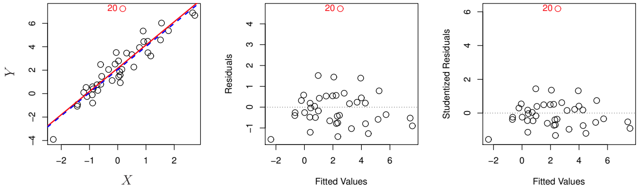


While they may not affect the fit, they might affect our assessment of model quality.

Possible solutions:

Outliers

Outliers are points with very high errors.



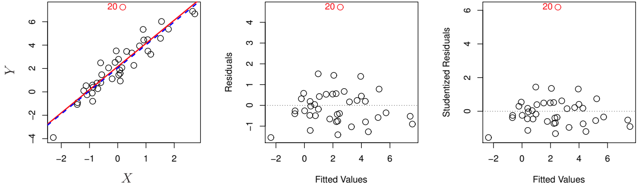
While they may not affect the fit, they might affect our assessment of model quality.

Possible solutions:

- If we believe an outlier is due to an error in data collection, we can remove it.

Outliers

Outliers are points with very high errors.



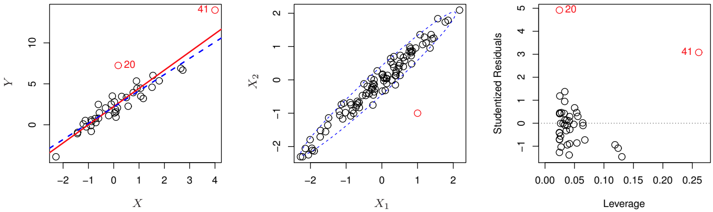
While they may not affect the fit, they might affect our assessment of model quality.

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.

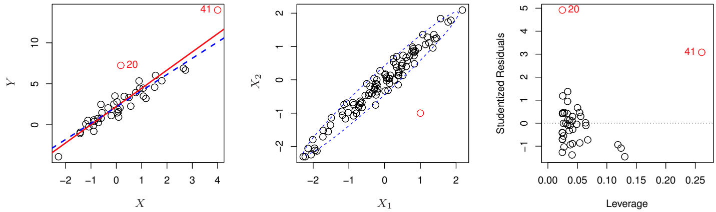
High leverage points

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



High leverage points

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

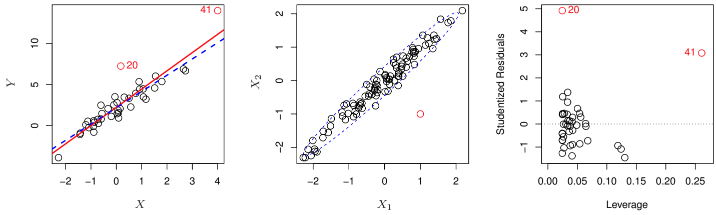


This can be measured with the **leverage statistic** or **self influence**:

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

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

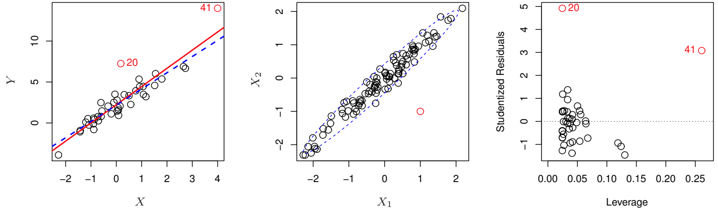
- The residual $\hat{\epsilon}_i = y_i - \hat{y}_i$ is an estimate for the noise ϵ_i.

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 $\sigma\sqrt{1 - h_{ii}}$.

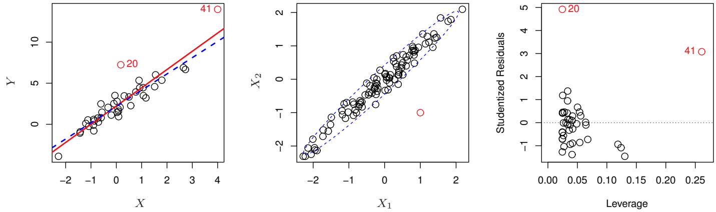
Studentized residuals

- 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.



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.
- When model is correct, it follows a Student-t distribution with $n-p-2$ degrees of freedom.

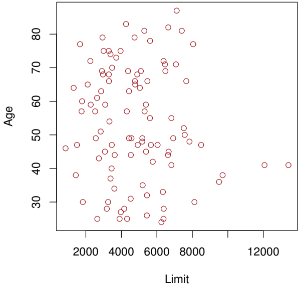


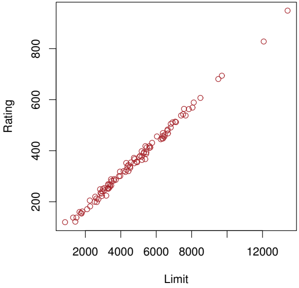
Collinearity

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`:

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

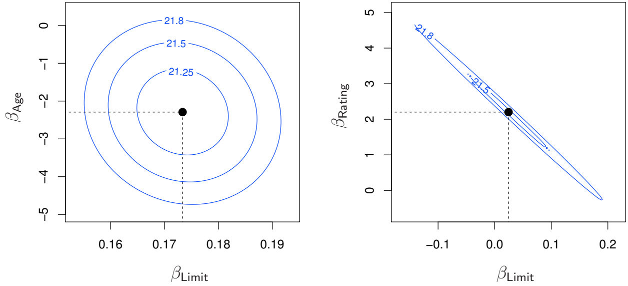
The fit $(\hat\beta_0, \hat\beta_1, \hat\beta_2)$ is just as good as $(\hat\beta_0, \hat\beta_1 + 100, \hat\beta_2 - 100)$.

Collinearity

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.

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.

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. **Inference**
KNN regression: prototypical nonparametric method. **Inference?**

Comparing Linear Regression to K-nearest neighbors

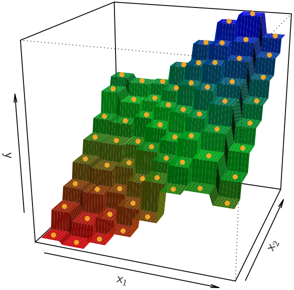
Linear regression: prototypical parametric method. **Inference**
KNN regression: prototypical nonparametric method. **Inference?**

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

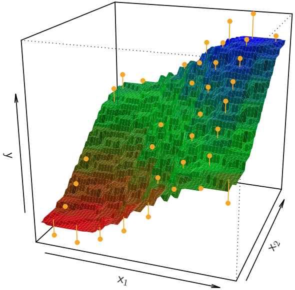
Comparing Linear Regression to K-nearest neighbors

Linear regression: prototypical parametric method. **Inference**
KNN regression: prototypical nonparametric method. **Inference?**

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



$K = 1$



$K = 9$

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

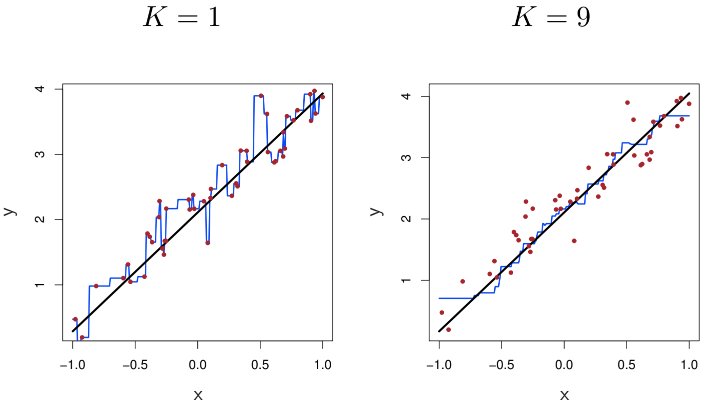
Linear regression: prototypical parametric method.

KNN regression: prototypical nonparametric method.

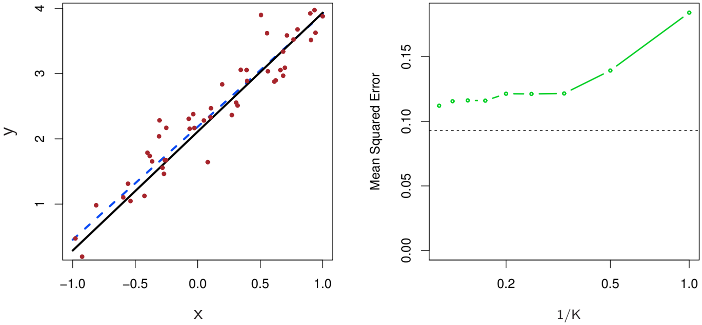
Long story short:

- 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.

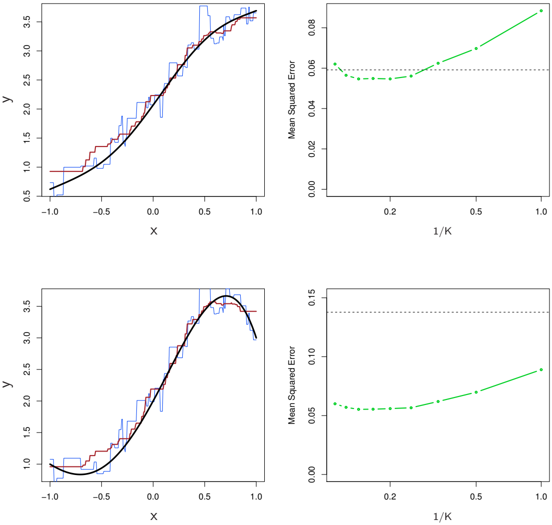
KNN estimates for a simulation from a linear model



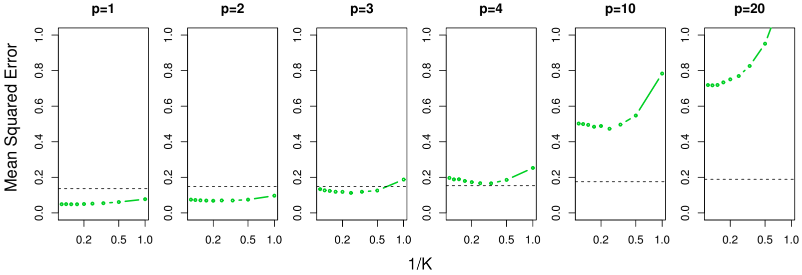
Linear models dominate KNN



Increasing deviations from linearity

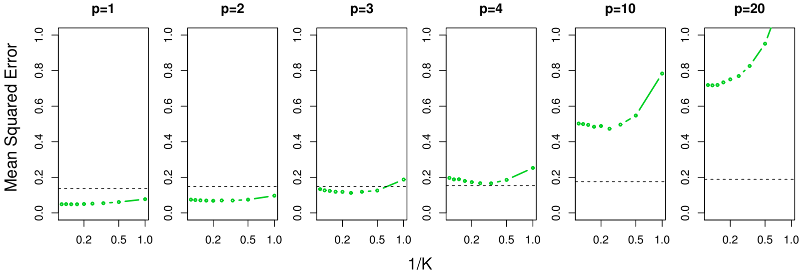


When there are more predictors than observations, Linear Regression dominates



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.