CKW

Statistics 8330: Data Analysis III
Linear Regression (review)

Suggested Reading: James, Witten, Hastie, and Tibshirani (JWHT), 2013, Chapter 3

Supplemental Reading: Hastie, Tibshirani, and Friedman (HTF), 2009, Chapter 3.1-3.3

Christopher K. Wikle

University of Missouri
Department of Statistics

Linear Regression

CKW

By this time in your statistics education, you have seen *linear regression* MANY times. Given that many of the more advanced statistical learning methods are based on variants of regression, it is worth doing a quick review, allowing us to emphasize certain points, some of which were brought up in the first lecture.

Given an output response variable Y, and an input vector $X^T = (X_1, X_2, \ldots, X_p)$, the linear regression model has the form

$$Y = f(X) + \epsilon,$$

where

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

It is important to note that we are specifying a model for the *conditional expectation*, $E(Y|X)$, and that we are *assuming* that this relationship is parametric and linear (or, at least reasonably approximated by a linear function). What does linear mean here?

Linear Regression Function

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j.$$

We assume that the β_j's are unknown parameters and that the variables (features) X_j can be:

- quantitative inputs
- transformations of quantitative inputs (e.g., square-root, log, etc.)
- polynomial terms: e.g., $X_2 = X_1^2$, $X_3 = X_1^3$, etc.
- interactions: e.g., $X_3 = X_1 \cdot X_2$, etc. (hierarchical principle)
- numeric or "dummy" coding of qualitative inputs. E.g., if G is a five-level factor input, then we might create $X_j, j = 1, \ldots, 4$ (assuming an intercept) such that $X_j = I(G = j)$ (note, $I(\cdot)$ is an indicator function that takes the value 1 when the argument is true and 0 otherwise). As we talked about in Stat 8310, there are different codings that could be used for qualitative variables - this effects the interpretation of parameters but not the overall fit or model statistics.

Fitting the Model

Given training data $(x_1, y_1) \cdots (x_n, y_n)$ where $x_i \equiv (x_{i1}, \ldots, x_{ip})^T$, we consider the *least squares* estimates. That is, we seek the β that minimize the residual sum of squares (RSS):

$$RSS(\beta) = \sum_{i=1}^{n}(y_i - f(x_i))^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

(HTF 2009, Fig 3.1)

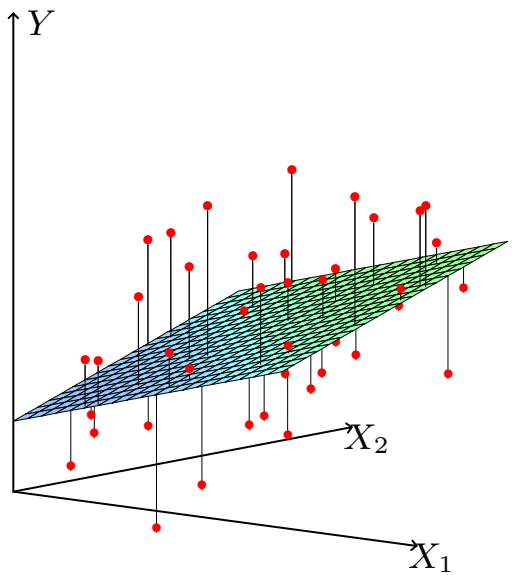


FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Fitting the Model: OLS

Let $\mathbf{X}$ be the $n \times (p+1)$ feature matrix (with an extra column for the intercept), and $\mathbf{y}$ the n-vector of responses, then we can write:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \equiv ||\mathbf{y} - \mathbf{X}\beta||^2$$

(i.e, the $L^2 norm$).

Then, differentiating $RSS(\beta)$ with respect to β and setting equal to zero we get the estimating ("normal") equations:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0,$$

which have the solution (when $\mathbf{X}$ has full column rank):

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

and

$$var(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2,$$

where $\sigma^2 = var(\epsilon)$.

OLS Prediction

Given an input vector x (assume it contains a 1 for the intercept), then fitted values are given by:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{H} \equiv \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the "hat matrix," which projects orthogonally the $\mathbf{y}$ vector onto the supspace occupied by $\mathbf{X}$: (HTF 2009, Fig 3.2)

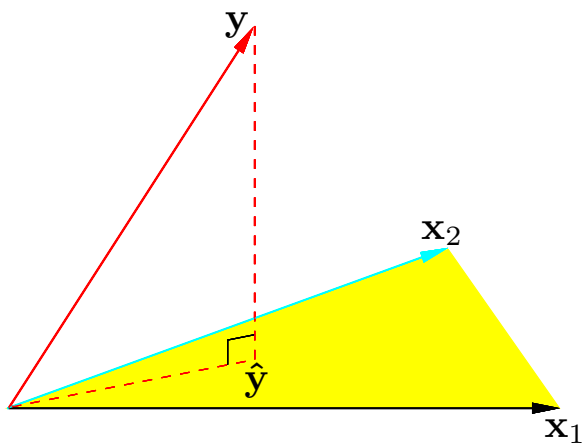


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

OLS Properties

CKW

A remarkable thing about the OLS estimates of parameters is that they are *unbiased* and have the smallest variance among *all* linear unbiased estimates! This is a consequence of the famous *Gauss-Markov Theorem* (see HTF 2009, 3.2.2). Note, in general, for an estimator $\tilde{\beta}$

$$MSE(\tilde{\beta}) = E(\tilde{\beta} - \beta)^2 = var(\tilde{\beta}) + [E(\tilde{\beta}) - \beta]^2 = \text{variance} + \text{bias}^2.$$

Thus, the second term is 0 for the OLS estimator ($\tilde{\beta} = \hat{\beta}$) since it is unbiased.

A major point that we will emphasize in this class is that when we are interested in prediction (especially), we care about MSE and we can sometimes get a much smaller MSE if we allow our estimator to be biased.

Thus, OLS estimates are not necessarily the "best" – the variance-bias tradeoff-is critical in statistical learning!

Inference

In practice, we estimate the variance by the unbiased estimator.

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Now, assuming that $\epsilon \sim N(0, \sigma^2)$, we get

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2),$$

and

$$(n-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1},$$

which allows inference on the model parameters and predictors (hypothesis tests, confidence intervals, prediction intervals).

Regression Inference

A typical question: *Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?* This corresponds to the overall hypothesis test:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

Recall, this is where we consider the F-statistic:

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

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Under the null hypothesis, this statistic has an F-distribution with p and $n - p - 1$ degrees of freedom.

Regression Inference (cont.)

More generally, given the null hypothesis when q coefficients are 0:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0,$$

the appropriate F-test constructed from the full and reduced model residual sums of squares:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)},$$

where RSS_0 is the residual sums of squares under the null assumptions, and the appropriate F-statistic has q and $n-p-1$ degrees of freedom.

This test corresponds to testing whether the q parameters are 0 given all of the other variables are in the model. In the case of looking at the partial effect of one variable, this test is equivalent to a partial t-test (i.e., tests the significance of a variable given all the others are in the model). These partial t-tests are reported in most regression summaries.

Example

CKW

Consider the advertising example in JWHT (2013, Chap. 3). Is there a significant linear relationship between Newspaper advertising and Sales?

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Simple regression of sales on newspaper

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	< 0.0001

Thus, although there is a relationship between Newspaper advertising and Sales, if TV and Radio are already in the model this relationship is not significant. Looking a little deeper shows that the Newspaper and Radio variables are fairly highly correlated with each other.

Inference (cont.)

Question: If you look at the results from a multiple regression and you find a significant partial t-statistic for one or more variables, do you still need to consider the overall F-test?

Yes. If p is fairly large, it is likely you will find at least one of the variables to be significant just by chance - thus, potentially assuming a relationship exists when it does not. The overall F-test does not suffer from this issue because it adjusts for the number of predictors.

Note: when $p > n$ (more features than samples) this all breaks down and we have to come up with other ways to do estimation and inference. This will be a major concern of ours in this class.

Regression: Model Fit

One can evaluate model fit by the coefficient of variation, R^2, or the root mean squared error (RMSE) [note, the JWHT book calls this residual standard error, RSE]. Recall,

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

and

$$RMSE = RSE = \sqrt{\frac{1}{n-p-1} RSS}.$$

Thus, although R^2 is a measure of the amount of variation that is removed from the response by the regression and falls between 0 and 1, it always increases as more variables are added. On the other hand, the RSE accounts for the number of variables in the model and so does not necessarily decrease as more variables are added. For future reference, recall that in traditional regression, we define the mean squared error to be, $MSE = RSS/(n-p-1)$.

Variable Selection

CKW

Often, we would like to know which of our predictors are associated with the response; i.e., *variable selection*. Recall from Stat 8310 that there are many measures we can use to compare models with different variables; e.g., AIC, BIC, Mallows C_p, R^2_{adj}, MSE. When possible, we would like to use a best-subset selection approach that can compare ALL models for a given number of predictors, p. But, when p is very large, this isn't possible. In that case, we might consider some sort of stepwise selection procedure:

- *forward selection* (start with no variables and add one by one)
- *backward selection* (start with all variables and remove one by one; note: can't be used if $p > n$)
- *mixed selection* (start with no variables and add – but, can remove one already added)

Issues

CKW

For real-world regression analyses one must evaluate the model for potential problems related to fit and violation of model assumptions. Some of the things we consider are:

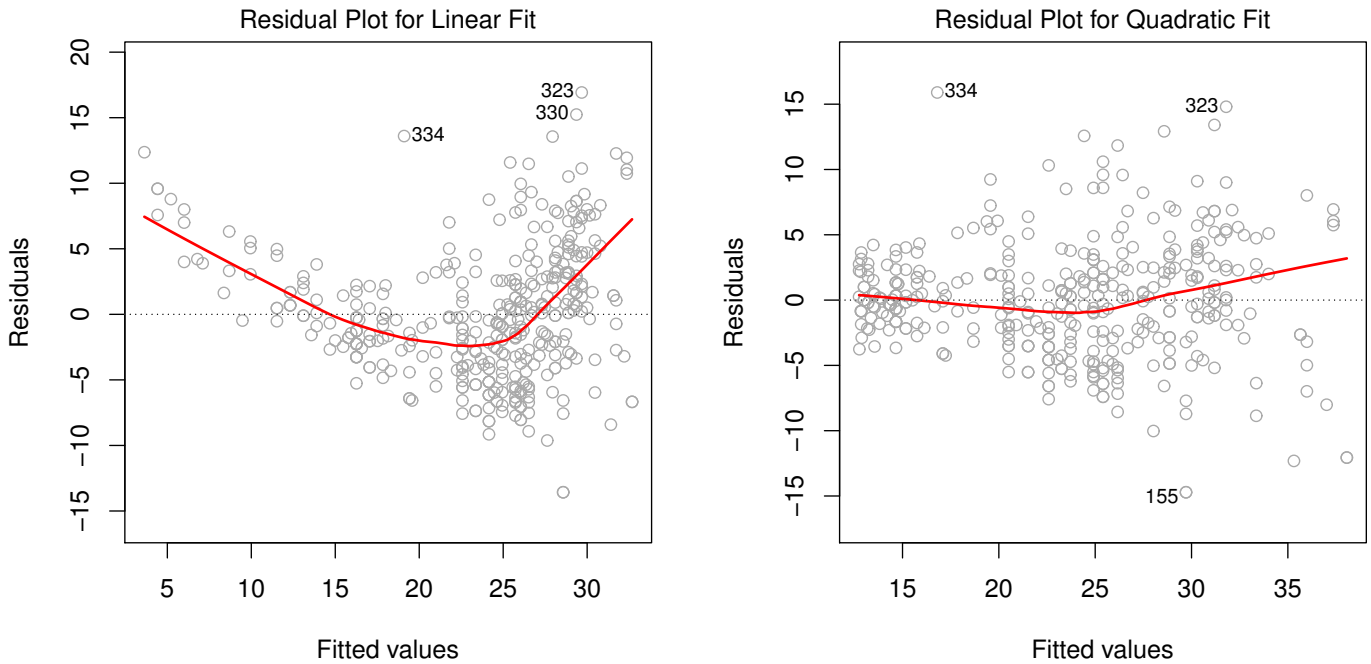
- Nonlinearity of the response-predictor relationship
- Correlated errors
- Non-constant error variance
- High-leverage points
- Collinearity

This can be a challenging task! Just because it is challenging is not a reason to ignore these issues. The more you do of it, the better you get at evaluating the assumptions.

Issues: Nonlinearity

If the true relationship is far from linear, then the inference is suspect and the model is not likely to produce good predictions.

We typically consider residual plots as a graphical tool to help identify these problems. In multiple regression, we typically plot the residuals versus the fitted values, $\hat{y}_i$. We hope to see no clear pattern in such a plot. Consider the example from JWHT (2013, Fig 3.9)



Issues: Correlated Errors

CKW

Recall we assume $\epsilon_i \sim$ *iid* $N(0, \sigma^2)$ - (i.e., independent errors); if they are not independent and we think they are, the estimated standard errors underestimate the true standard errors (why?)

Often, when data are measured in time and/or space, there are unaccounted for predictors that can lead to dependent errors (observations nearby in time and/or space tend to be more alike - thus, dependent). One can often see this by plotting residuals in order (for time) or on a map (in space), and look for distinct patterns or trends.

Issues: Correlated Errors

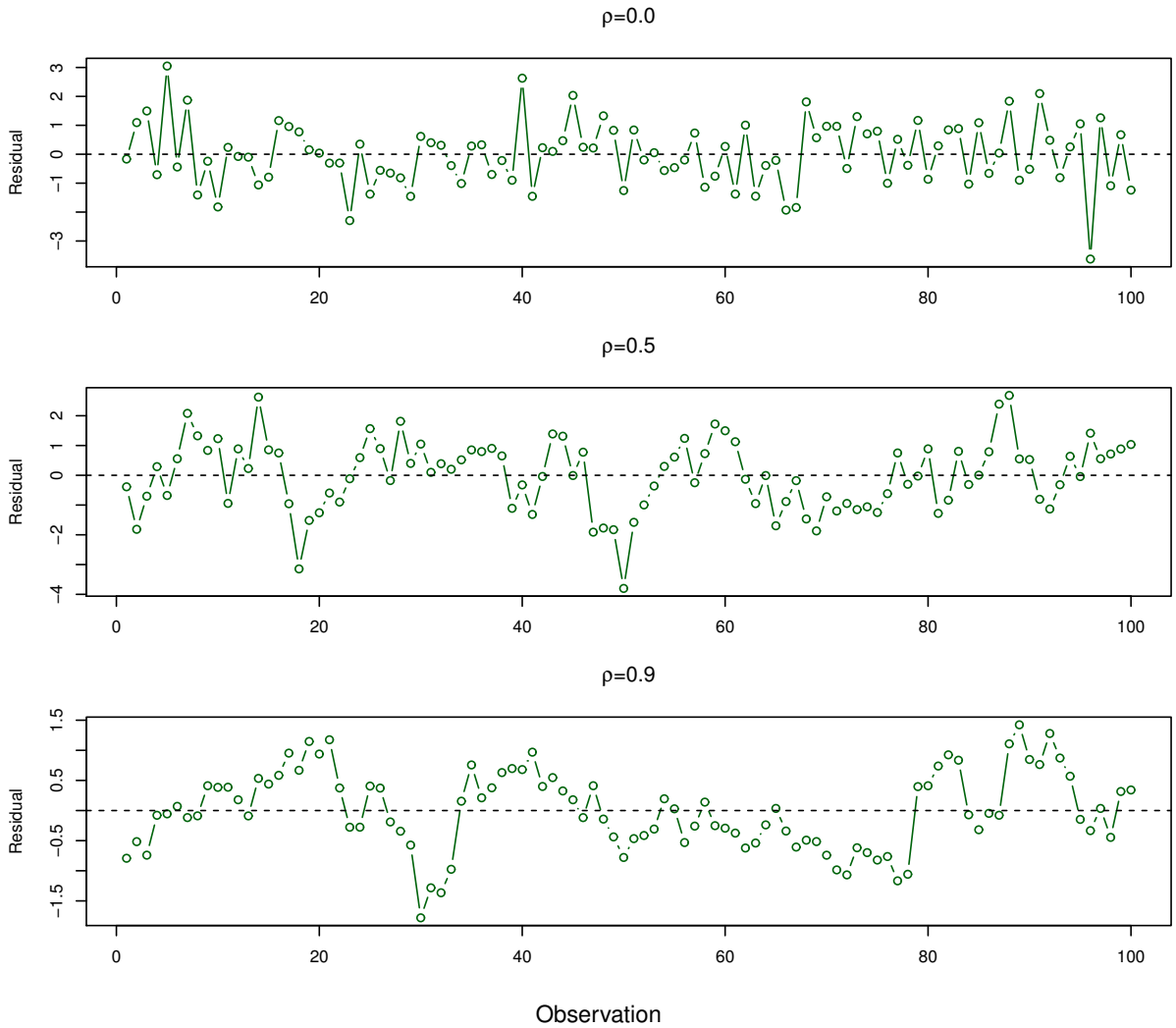


Figure 3.10 in JWHT (2013): Plots of residuals from simulated time series data sets generated with differing levels of correlation between error terms for adjacent time points.

Issues: Non-constant Error Variance

Recall, we assume $var(\epsilon_i) = \sigma^2$ for all i. Our inference and prediction intervals are sensitive to this assumption. Non-constant variances (*heteroscedasticity*) can be identified by plotting residuals against the fitted values. It can often be mitigated by transforming the response (e.g., log, square root) or using weighted least squares. Consider the example:

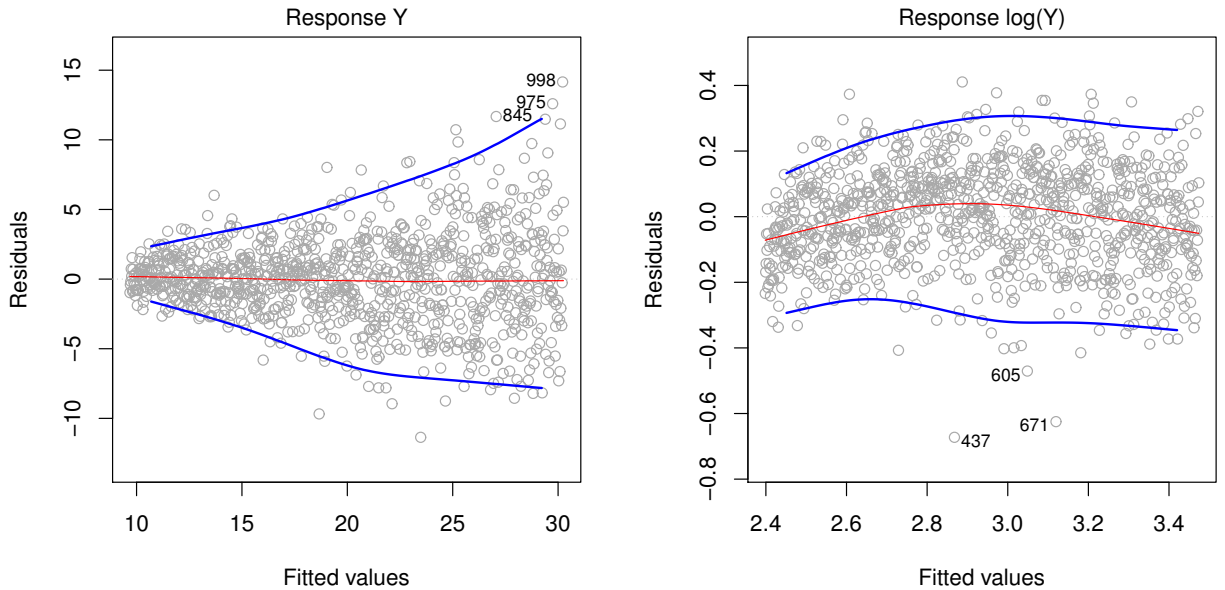
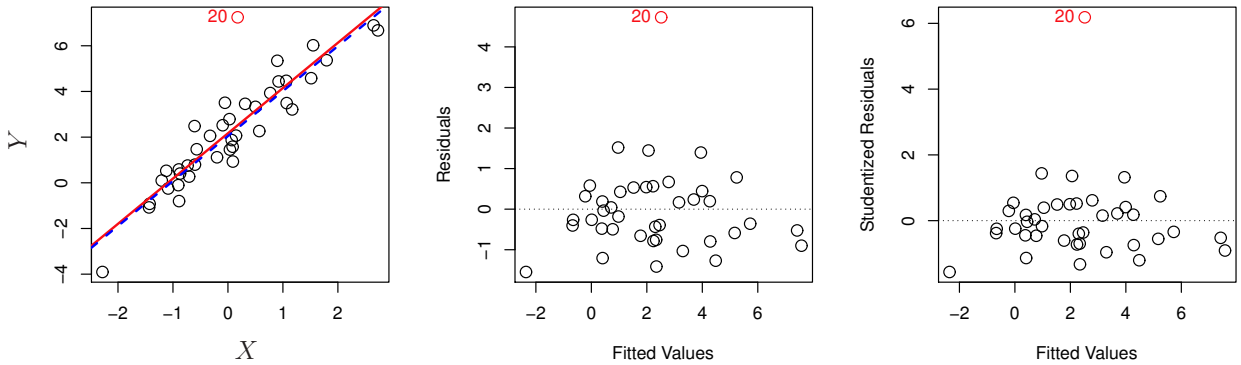


Figure 3.11 from JWHT (2013): Residual plots. Red line is smooth fit to the residuals, intended to make it easier to identify a trend. Blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: Funnel shape indicates heteroscedasticity. Right: Response has been log transformed.

Issues: Outliers

CKW

An outlier is simply a point for which y_i is far from the value predicted by the model. Often an outlier doesn't have a strong effect on the fit of the model, but it can still have negative consequences because increasing the RSE inflates our estimate $\hat{\sigma}^2$ and thus, our confidence intervals. We can often see these by plotting residuals or studentized residuals. There are also several diagnostic statistics that can be used (see Stat 8310 notes). Consider the example from JWHT (2013), Figure 3.12:

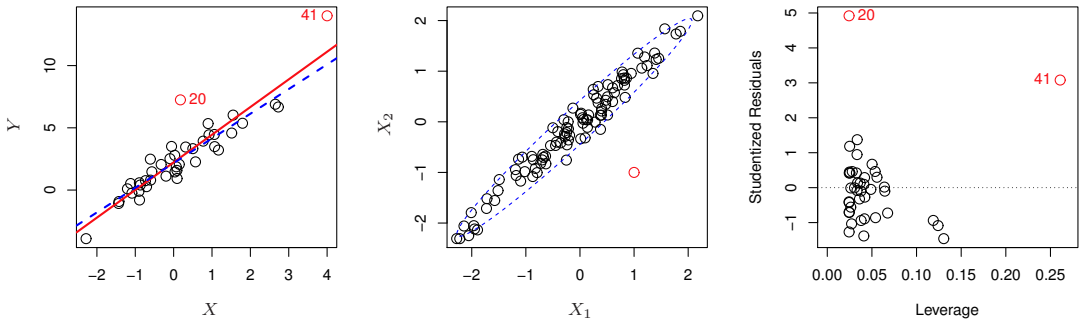


Left: Least squares regression line in red, and the regression line after removing the outlier in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual of 6; typically we expect values between -3 and 3.

Issues: High Leverage Points

CKW

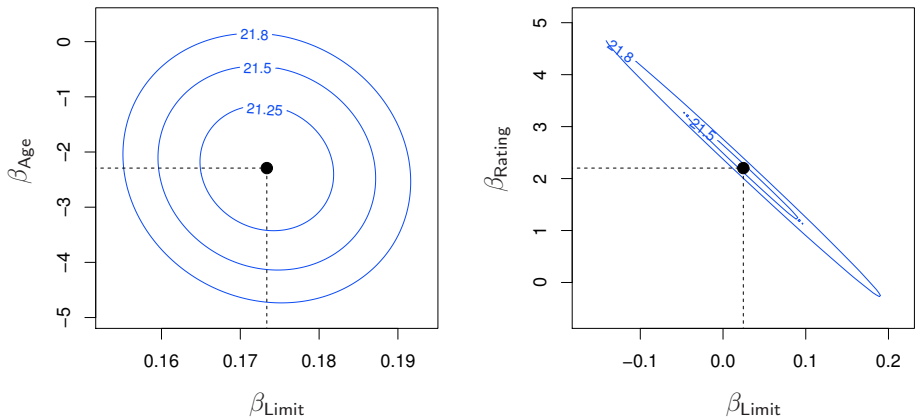
High leverage points are observations that have unusual x values such that they have too much influence on the regression fit. Recall that we can identify high leverage points by looking at the diagonal of the hat matrix, $\mathbf{H}$. One rule of thumb says that if the hat diagonal element (h_{ii}) greatly exceeds $(p+1)/n$, then the point is likely to have high leverage. Consider Figure 3.14 from JWHT (2013).



Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X1 value or its X2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.

Issues: Collinearity

Multicollinearity refers to the situation where two or more predictor variables are closely (linearly) related. In many respects, collinearity is the most insidious of all of the regression issues because it is almost ALWAYS present in real-world datasets with large p and it does not necessarily show itself unless you look for it. The problem is that the fitted regression coefficients can change dramatically when collinearity is present – this can seriously affect inference (it doesn't affect predictions as much). Consider Figure 3.15 from JWHT (2013).



Contour plots of RSS values as a function of the βs. Black dots represent the coefficient values for the minimum RSS. Left: A contour plot of RSS for the regression of balance onto age and limit. Minimum value is well defined. Right: A contour plot of RSS for the regression of balance onto rating and limit. Because of the collinearity, many pairs with a similar value for RSS.

Issues: Collinearity (cont.)

CKW

In addition to detecting collinearity from predictor correlation estimates, dramatic changes in regression coefficient estimates given new variables, larger than expected standard errors, and unexpected signs on regression estimates, one can identify collinearity by condition numbers and variance inflation factors (VIFs). Recall,

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

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors (X_{-j}). When this coefficient of variation is close to 1, the associated VIF is very large. Thus, we look for large VIF values to identify multicollinearity.

Parametric Regression?

CKW

Parametric multiple regression models are, in many respects, the default model in statistics. Why?

Multiple regression models are *parametric* because they assign a particular linear functional form for $f(X)$. This is advantageous for several reasons:

- They are easy to fit
- They have a relatively small number of parameters to estimate
- Coefficients have simple interpretations (when collinearity isn't present)
- Statistical tests are easy to perform

But, linear regressions have a pretty serious disadvantage as well:

- They make a VERY strong assumption about the form of $f(X)$; if this functional form is wrong, then inference will be inappropriate and predictions will not be good

A Non-parametric Alternative

CKW

In contrast to parametric regression models, *non-parametric* methods do not assume explicitly a parametric form for $f(X)$. The advantage of this is that they are more flexible and can accommodate more complicated relationships between the predictors and the response.

Let's consider perhaps the simplest non-parametric regression model, the *K-nearest neighbors regression* (KNN regression). This is similar to the KNN classifier we talked about last lecture.

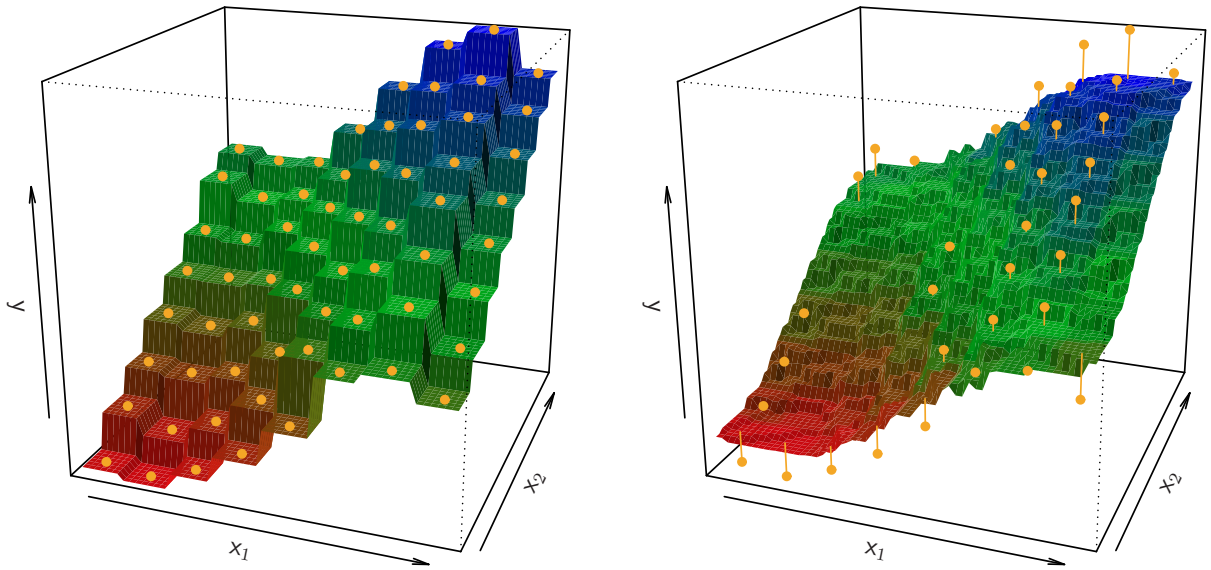
Given a value for K (number of neighbors) and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, which we call $\mathcal{N}_0$. Then, $f(x_0)$ is estimated using the average of all of the training *responses* in $\mathcal{N}_0$:

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i.$$

KNN Regression (cont.)

CKW

Consider the example in JWHT (2013) Figure 3.16.



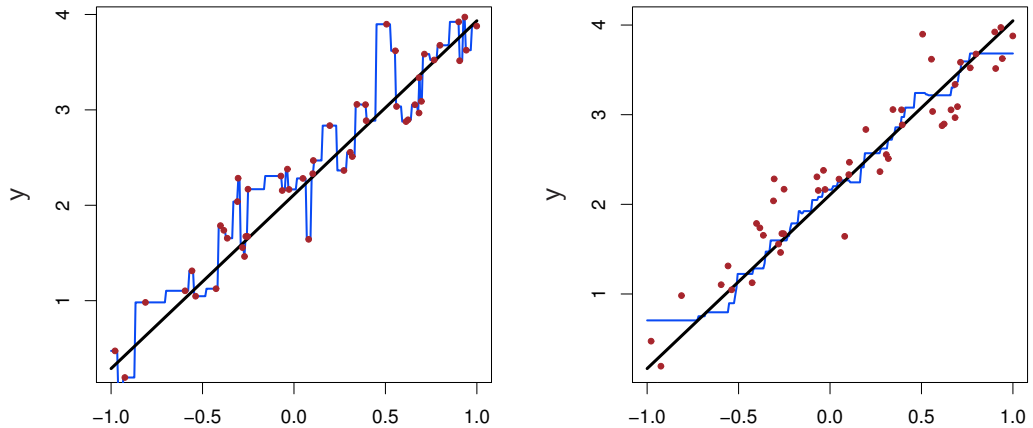
This plot illustrates one of the most important issues we face when fitting nonparametric models – the *bias-variance tradeoff*. Note, when $K = 1$ the fit is perfect for the training observations (no bias) but quite rough (more variance). In contrast, when $K = 9$ the fit is much smoother (less variance) but does not go through the data points (more bias).

Non-parametric vs Parametric

CKW

How do parametric and non-parametric regressions compare (other than interpretation, ease of fit, etc. mentioned earlier)?

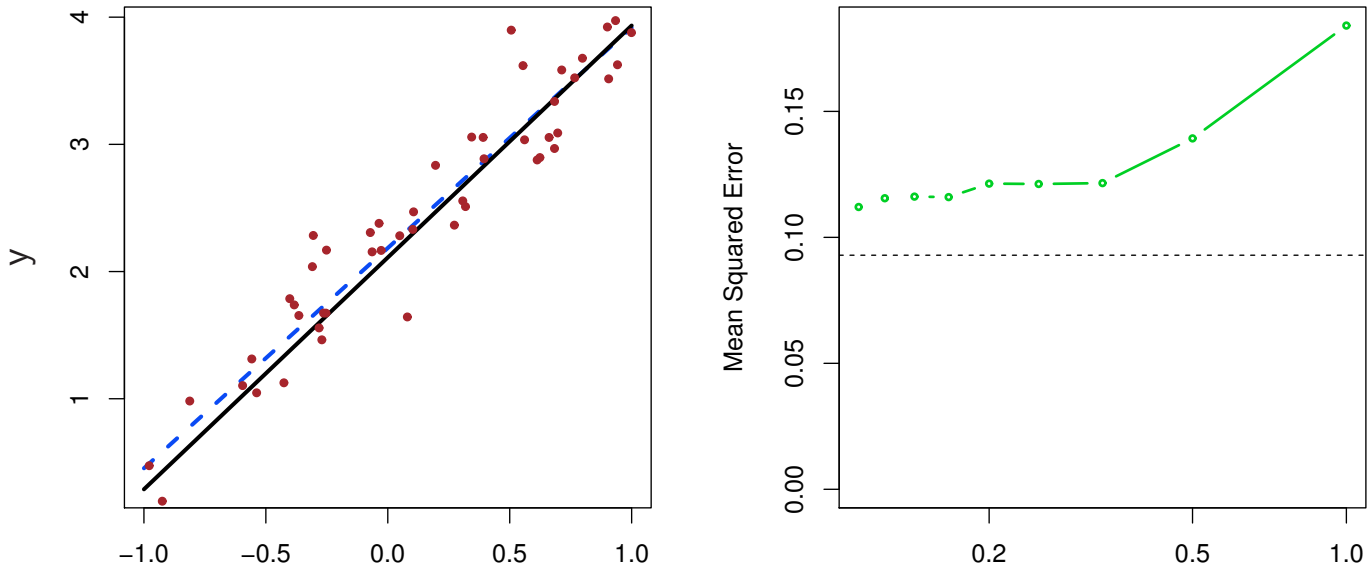
The parametric least squares approach will outperform the non-parametric approach if the parametric form for $f(X)$ that has been selected is close to the true form of $f(X)$. Consider the examples from JWHT (2013). E.g., Figure 3.17:



Plots of f(X) using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line. Left: The blue curve corresponds to K = 1 and interpolates (i.e. passes directly through) the training data. Right: The blue curve corresponds to K = 9, and represents a smoother fit.

OLS vs KNN (cont.)

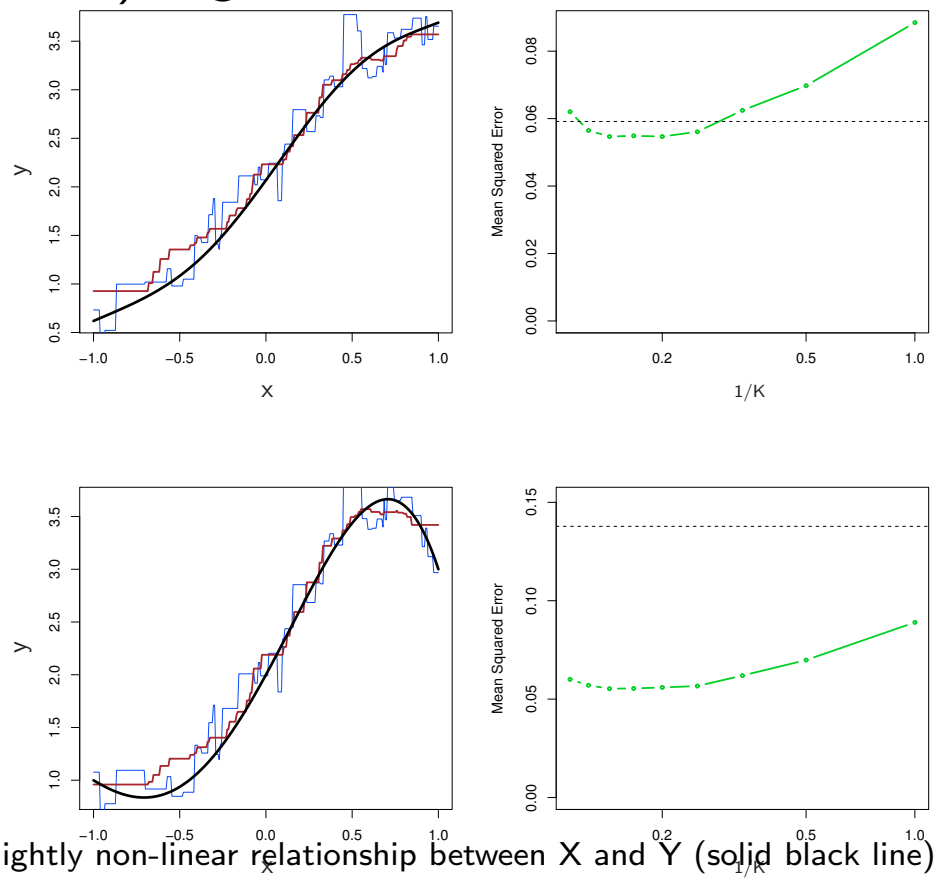
Consider JWHT (2013) Figure 3.18:



The same data set shown in Figure 3.17 is investigated further. Left: The blue dashed line is the least squares fit to the data. Since f(X) is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of f(X). Right: The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of 1/K (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since f(X) is in fact linear. For KNN regression, the best results occur with a very large value of K, corresponding to a small value of 1/K.

OLS vs KNN (cont.)

Consider JWHT (2013) Figure 3.19:



Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with K = 1 (blue) and K = 9 (red) are displayed. Top Right: For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of 1/K (green) are displayed. Bottom Left and Bottom Right: As in the top panel, but with a strongly non-linear relationship between X and Y .

OLS vs KNN (cont.)

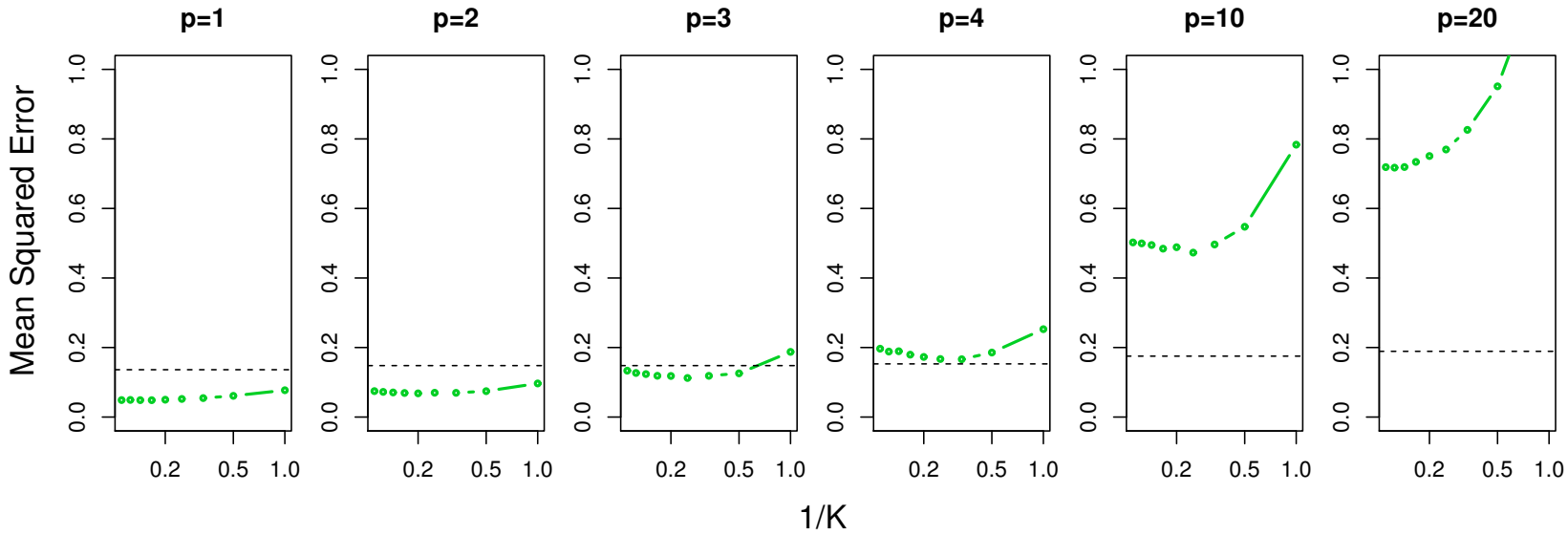
Importantly, although we have seen that KNN regression can out-perform OLS regression when the regression function is misspecified, there is a serious limitation to KNN related to the curse-of-dimensionality discussed in the last lecture.

Non-parametric methods require a lot of samples to be able to identify the flexible response function.

When p is large, there are not enough observations to be able to do this (neighbors are difficult to come by!), and the methods perform poorly.

OLS vs KNN (cont.)

As a general rule, parametric methods will tend to outperform non-parametric methods when the number of observations per predictor is small. Consider Figure 3.20 from JWHT (2013):



Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non-linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.