

Lecture 3: Linear Regression

STATS 202: Data Mining and Analysis

Linh Tran

tranlm@stanford.edu



Department of Statistics
Stanford University

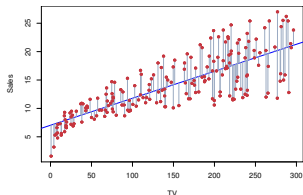
June 28, 2021



- ▶ Reference textbook for statistics
 - ▶ *Grinstead and Snell*
- ▶ HW1 due in one week (no class).
- ▶ Section on R/Python programming for DS this Friday.
- ▶ Please enroll in Piazza/Gradescope.



- ▶ Linear regression
 - ▶ Coefficients, standard errors, hypothesis testing
- ▶ Multiple linear regression
 - ▶ Variable selection, stepwise models, categorical variables,
- ▶ Regression issues
 - ▶ Interactions, non-linear relationships, error correlation, heteroskedasticity



Example of a linear model fit to some data.

Recall:

- ▶ Given some input features X_1, X_2, \dots, X_p
- ▶ $Y \in \mathbb{R}$ is the output
- ▶ (X, Y) have a joint distribution
- ▶ **Blue line** is the regression fit: an estimate \hat{f}_n of the line we want

$$f_0 = \mathbb{E}_0[Y|X_1, X_2, \dots, X_p] \quad (1)$$

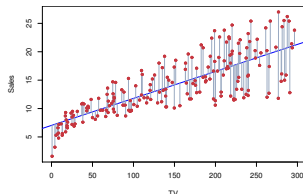


In linear regression, we assume

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (2)$$

$$\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad (3)$$

$$\mathbb{E}[y|x] = \beta_0 + \beta_1 x \quad (4)$$



Example of a linear model fit to some data.

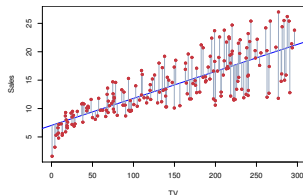


In linear regression, we assume

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (2)$$

$$\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad (3)$$

$$\mathbb{E}[y|x] = \beta_0 + \beta_1 x \quad (4)$$



Example of a linear model fit to some data.

We can get coefficient estimates $(\hat{\beta}_0, \hat{\beta}_1)$ by minimizing some objective function, e.g. the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (5)$$

$$= \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad (6)$$



Some calculus shows that the minimizers of the RSS are:

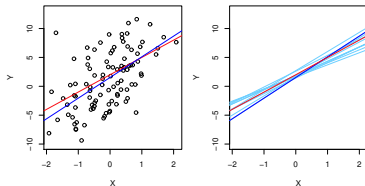
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (7)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (8)$$

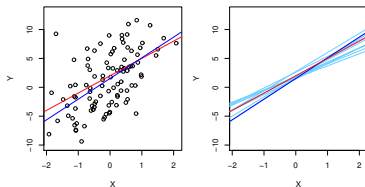
where \bar{y} and \bar{x} are the sample averages of y_i and x_i , respectively.



- ▶ Different samples will result in different estimates $(\hat{\beta}_0, \hat{\beta}_1)$
- ▶ How do we evaluate the certainty of $(\hat{\beta}_0, \hat{\beta}_1)$?



True function f_0 and
estimate \hat{f}_n .

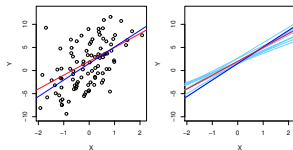


True function f_0 and
estimate \hat{f}_n .

- ▶ Different samples will result in different estimates $(\hat{\beta}_0, \hat{\beta}_1)$
- ▶ How do we evaluate the certainty of $(\hat{\beta}_0, \hat{\beta}_1)$?
- ▶ **Recall:** When estimating mean μ_0 of variable X , we can compute its standard error $SE(\hat{\mu}_n)$ as

$$SE(\hat{\mu}_n) = \sqrt{\frac{\sigma_0^2}{n}} \quad (9)$$

- ▶ We can take a similar approach with our coefficients
 - ▶ i.e. estimate standard errors

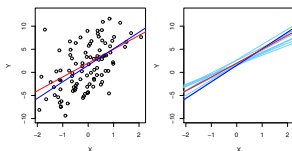


True function f_0 and estimates \hat{f}_n .

$$\begin{aligned} SE(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right] \\ SE(\hat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned} \quad (10)$$

where $\sigma^2 = \text{Var}(\epsilon)$.

- Assumes ϵ_i are uncorrelated with common variance σ^2

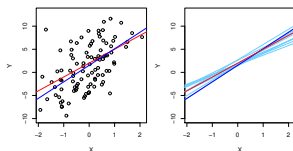


True function f_0 and estimates \hat{f}_n .

- While, we don't know σ_0 , we can estimate it

$$\begin{aligned}\hat{SE}(\hat{\beta}_0)^2 &= \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right] \\ \hat{SE}(\hat{\beta}_1)^2 &= \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}\tag{11}$$

where $\hat{\sigma} = \sqrt{RSS/(n-2)}$.



True function f_0 and estimates \hat{f}_n .

- While, we don't know σ_0 , we can estimate it

$$\begin{aligned}\hat{SE}(\hat{\beta}_0)^2 &= \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right] \\ \hat{SE}(\hat{\beta}_1)^2 &= \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}\quad (11)$$

where $\hat{\sigma} = \sqrt{RSS/(n-2)}$.
95% CI's can then be calculated:

$$\hat{\beta}_0 \pm t_{\alpha/2} \cdot \hat{SE}(\hat{\beta}_0) \quad (12)$$

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot \hat{SE}(\hat{\beta}_1) \quad (13)$$



When we want to evaluate some kind of relationship, we can test it statistically, e.g.

$$H_0 : \text{There is no relationship between } X \text{ and } Y \quad (14)$$

$$H_a : \text{There is a relationship between } X \text{ and } Y \quad (15)$$



When we want to evaluate some kind of relationship, we can test it statistically, e.g.

$$H_0 : \text{There is no relationship between } X \text{ and } Y \quad (14)$$

$$H_a : \text{There is a relationship between } X \text{ and } Y \quad (15)$$

Note: Hypothesis tests are typically set up such that H_a is the outcome that we care about

- ▶ e.g. In non-inferiority tests, H_0 is typically specified such that there **is** a deficiency in the treatment being evaluated.



For linear models, we typically test e.g.

$$H_0 : \beta_1 = 0 \quad (16)$$

$$H_a : \beta_1 \neq 0 \quad (17)$$

- If $\beta_1 = 0$, then our model simplifies to $\mathbb{E}[y|x] = \beta_0$, meaning X is not associated to Y .



For linear models, we typically test e.g.

$$H_0 : \beta_1 = 0 \quad (16)$$

$$H_a : \beta_1 \neq 0 \quad (17)$$

- ▶ If $\beta_1 = 0$, then our model simplifies to $\mathbb{E}[y|x] = \beta_0$, meaning X is not associated to Y .
- ▶ To be sure $\beta_1 \neq 0$, we want $\hat{\beta}_1$ to be far from 0 and for $\hat{SE}(\hat{\beta}_1)$
- ▶ Will typically calculate a statistic to help us evaluate this
 - ▶ e.g. A t -statistic



For linear models, we typically test e.g.

$$H_0 : \beta_1 = 0 \quad (18)$$

$$H_a : \beta_1 \neq 0 \quad (19)$$

Our test statistic

$$t = \frac{\hat{\beta}_1 - 0}{\hat{SE}(\hat{\beta}_1)} \quad (20)$$



For linear models, we typically test e.g.

$$H_0 : \beta_1 = 0 \quad (18)$$

$$H_a : \beta_1 \neq 0 \quad (19)$$

Our test statistic

$$t = \frac{\hat{\beta}_1 - 0}{\hat{SE}(\hat{\beta}_1)} \quad (20)$$

- ▶ Follows a t-distribution with $n - 2$ degrees of freedom.
- ▶ Can be used to calculate a p-value
 - ▶ i.e. the probability of observing our statistic (or a larger one) under the null hypothesis
 - ▶ If the probability is low enough, then we reject H_0



An applied example

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. For the **Advertising** data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of \$1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the **sales** variable is in thousands of units, and the **TV** variable is in thousands of dollars).



1. If we reject the null hypothesis, can we assume there is a linear relationship?



1. If we reject the null hypothesis, can we assume there is a linear relationship?
 - ▶ No. A quadratic relationship may be a better fit, for example.



1. If we reject the null hypothesis, can we assume there is a linear relationship?
 - ▶ No. A quadratic relationship may be a better fit, for example.
2. If we don't reject the null hypothesis, can we assume there is no relationship between X and Y ?



1. If we reject the null hypothesis, can we assume there is a linear relationship?
 - ▶ No. A quadratic relationship may be a better fit, for example.
2. If we don't reject the null hypothesis, can we assume there is no relationship between X and Y ?
 - ▶ No. This test is only powerful against certain monotone alternatives (with enough data). There could be more complex non-linear relationships (or you could need more data).



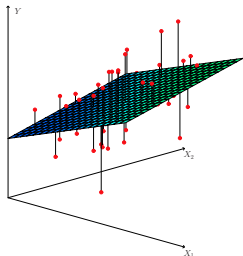
Extension of linear regression to handle multiple predictors

In multiple linear regression, we assume

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$$

$$\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad (21)$$





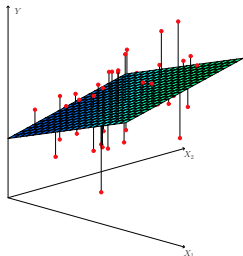
Extension of linear regression to handle multiple predictors

In multiple linear regression, we assume

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon$$

$$\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\mathbb{E}[Y|\mathbf{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad (21)$$



In matrix notation:

$$\mathbb{E}[Y|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta} \quad (22)$$

where

$$\mathbf{X} = (1, X_1, X_2, \dots, X_p) \quad (23)$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top \quad (24)$$



- ▶ Is at least one of the variables X_j useful for predicting the outcome Y ?



- ▶ Is at least one of the variables X_j useful for predicting the outcome Y ?
- ▶ Which subset of the predictors is most important?



- ▶ Is at least one of the variables X_j useful for predicting the outcome Y ?
- ▶ Which subset of the predictors is most important?
- ▶ How good is a linear model for these data?



- ▶ Is at least one of the variables X_j useful for predicting the outcome Y ?
- ▶ Which subset of the predictors is most important?
- ▶ How good is a linear model for these data?
- ▶ Given a set of predictor values, what is a likely value for Y , and how accurate is this prediction?



Our goal is the same: minimize the RSS

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (25)$$

$$= \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p}))^2 \quad (26)$$

Can be show that RSS is miminized with:

$$\beta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (27)$$

where the vectors are now matrices, e.g.

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix} \quad (28)$$



Our goal is the same: minimize the RSS

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (25)$$

$$= \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_p x_{i,p}))^2 \quad (26)$$

Can be show that RSS is mimimized with:

$$\beta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (27)$$

where the vectors are now matrices, e.g.

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p} \end{bmatrix} \quad (28)$$

Note: only exists when $\mathbf{X}^\top \mathbf{X}$ is invertible (requires $n \geq p$).

Which variables are important?



Consider the hypothesis:

H_0 : The last q predictors have no relation with Y . (29)

Which variables are important?



Consider the hypothesis:

H_0 : The last q predictors have no relation with Y . (29)

i.e. H_0 : $\beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$ (30)

Which variables are important?



Consider the hypothesis:

H_0 : The last q predictors have no relation with Y . (29)

i.e. H_0 : $\beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$ (30)

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F -statistic is defined by:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)} \quad (31)$$

Under the null hypothesis, statistic follows F -distribution.

Which variables are important?



Consider the hypothesis:

H_0 : The last q predictors have no relation with Y . (29)

i.e. H_0 : $\beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$ (30)

Let RSS_0 be the residual sum of squares for the model which excludes these variables. The F -statistic is defined by:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)} \quad (31)$$

Under the null hypothesis, statistic follows F -distribution.

Example: If $q = p$, testing if $\beta_j = 0 \forall j$.

$$RSS_0 = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (32)$$

Which variables are important?



Some notes:

- ▶ The t -statistic associated to the j^{th} predictor is (equivalent to) the square root of the F -statistic for the null hypothesis which sets only $\beta_j = 0$.



Some notes:

- ▶ The t -statistic associated to the j^{th} predictor is (equivalent to) the square root of the F -statistic for the null hypothesis which sets only $\beta_j = 0$.
- ▶ A low p -value for the j^{th} predictor indicates that the predictor is important.



Some notes:

- ▶ The t -statistic associated to the j^{th} predictor is (equivalent to) the square root of the F -statistic for the null hypothesis which sets only $\beta_j = 0$.
- ▶ A low p -value for the j^{th} predictor indicates that the predictor is important.
- ▶ **Warning:** If there are many predictors, even under the null hypothesis, some of the t -tests will have low p -values.



Some notes:

- ▶ The t -statistic associated to the j^{th} predictor is (equivalent to) the square root of the F -statistic for the null hypothesis which sets only $\beta_j = 0$.
- ▶ A low p -value for the j^{th} predictor indicates that the predictor is important.
- ▶ **Warning:** If there are many predictors, even under the null hypothesis, some of the t -tests will have low p -values. Ways of accounting for this include e.g.
 - ▶ controlling the family-wise error rate (FWER)
 - ▶ controlling the false discovery rate (FDR)

Which variables are important?



Example of multiple linear regression output (in R):

```
Residuals:
    Min       1Q   Median       3Q      Max
-15.594   -2.730   -0.518    1.777   26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim         -1.080e-01  3.286e-02  -3.287 0.001087 **
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288
chas         2.687e+00  8.616e-01   3.118 0.001925 **
nox          -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116 < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229
dis          -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad           3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax          -1.233e-02  3.761e-03  -3.280 0.001112 **
ptratio      -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat        -5.248e-01  5.072e-02  -10.347 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-Squared:  0.7406,    Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```


How many variables are important?



In selecting a subset of the predictors, we have 2^p choices.

One way to simplify the choice is to define a range of models with an increasing number of variables, then select the best. AKA stepwise regression.

The approach:

1. Construct a sequence of p models with increasing number of variables.
2. Select the best model among them.

How many variables are important?



Constructing the p models:

- ▶ *Forward selection*: Starting from a *null* model, include variables one at a time, minimizing the RSS at each step.



Constructing the p models:

- ▶ *Forward selection*: Starting from a *null* model, include variables one at a time, minimizing the RSS at each step.
- ▶ *Backward selection*: Starting from the *full* model, eliminate variables one at a time, choosing the one with the largest p -value at each step.



Constructing the p models:

- ▶ *Forward selection*: Starting from a *null* model, include variables one at a time, minimizing the RSS at each step.
- ▶ *Backward selection*: Starting from the *full* model, eliminate variables one at a time, choosing the one with the largest p -value at each step.
- ▶ *Mixed selection*: Starting from a *null* model, include variables one at a time, minimizing the RSS at each step. If the p -value for some variable goes beyond a threshold, eliminate that variable.



Constructing the p models:

- ▶ *Forward selection*: Starting from a *null* model, include variables one at a time, minimizing the RSS at each step.
- ▶ *Backward selection*: Starting from the *full* model, eliminate variables one at a time, choosing the one with the largest p -value at each step.
- ▶ *Mixed selection*: Starting from a *null* model, include variables one at a time, minimizing the RSS at each step. If the p -value for some variable goes beyond a threshold, eliminate that variable.

Choosing a model in the range produced is a form of tuning. Will cover this more in Chapter 6.

How many variables are important?



Example output of a stepwise selection method:

- ▶ {}
- ▶ {tv}
- ▶ {tv, newspaper}
- ▶ {tv, newspaper, radio}
- ▶ {tv, newspaper, radio, facebook}
- ▶ {tv, newspaper, radio, facebook, twitter}

6 choices are better than $2^6 = 64$.

We can use different objectives to decide on optimal model, e.g. cross-validation, AIC, BIC, etc.

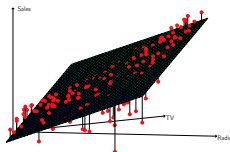


To assess fit, we focus on the residuals.

- ▶ The RSS always decreases as we add more variables.
- ▶ The residual standard error (RSE) corrects this:

$$RSE = \sqrt{\frac{1}{n - p - 1} RSS} \quad (33)$$

- ▶ Visualizing the residuals can reveal phenomena that are not accounted for by the model; eg. synergies or interactions:



How good is the predictions?



We can get confidence intervals for our predictions:

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="confidence")  
      fit    lwr    upr  
1 29.80 29.01 30.60  
2 25.05 24.47 25.63  
3 20.30 19.73 20.87
```

The confidence intervals reflect the uncertainty from $\hat{\beta}$.

How good is the predictions?



We can get confidence intervals for our predictions:

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="confidence")  
      fit    lwr    upr  
1 29.80 29.01 30.60  
2 25.05 24.47 25.63  
3 20.30 19.73 20.87
```

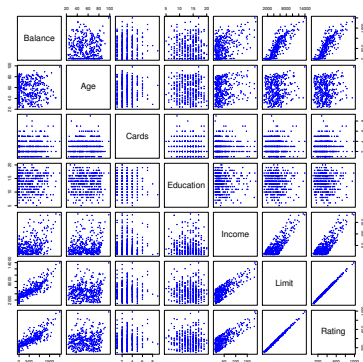
The confidence intervals reflect the uncertainty from $\hat{\beta}$.

```
> predict(lm.fit, data.frame(lstat=(c(5,10,15))),  
          interval="prediction")  
      fit    lwr    upr  
1 29.80 17.566 42.04  
2 25.05 12.828 37.28  
3 20.30  8.078 32.53
```

Prediction intervals reflect uncertainty from **both** $\hat{\beta}$ and ϵ (i.e. the irreducible error).



Example: credit dataset



Additionally:

4 qualitative variables

- ▶ gender: male, female
- ▶ student: yes, no
- ▶ status: married, single, divorced
- ▶ ethnicity: African American, Asian, Caucasian

Example of a linear model fit to some data.



For each qualitative predictor, e.g. ethnicity:

- ▶ Choose a baseline category, e.g. African American
 - ▶ Can be the group with the highest frequency



For each qualitative predictor, e.g. ethnicity:

- ▶ Choose a baseline category, e.g. African American
 - ▶ Can be the group with the highest frequency
- ▶ For every other category, define a new predictor (aka dummy indicator):
 - ▶ X_{Asian} is 1 if the person is Asian and 0 otherwise.
 - ▶ $X_{Caucasian}$ is 1 if the person is Caucasian and 0 otherwise.



For each qualitative predictor, e.g. ethnicity:

- ▶ Choose a baseline category, e.g. African American
 - ▶ Can be the group with the highest frequency
- ▶ For every other category, define a new predictor (aka dummy indicator):
 - ▶ X_{Asian} is 1 if the person is Asian and 0 otherwise.
 - ▶ $X_{Caucasian}$ is 1 if the person is Caucasian and 0 otherwise.
- ▶ The model will be:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{Asian} X_{Asian} + \beta_{Caucasian} X_{Caucasian} + \epsilon \quad (34)$$

β_{Asian} is the relative effect on balance for being Asian compared to the baseline category.



$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{Asian} X_{Asian} + \beta_{Caucasian} X_{Caucasian} + \epsilon \quad (35)$$

- ▶ The model fit and predictions are independent of the choice of the baseline category.
- ▶ Other ways to encode qualitative predictors produce the same fit \hat{f}_n , but the coefficients have different interpretations.
- ▶ Hypothesis tests derived from these dummy indicator are affected by the choice.



$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{Asian} X_{Asian} + \beta_{Caucasian} X_{Caucasian} + \epsilon \quad (35)$$

- ▶ The model fit and predictions are independent of the choice of the baseline category.
- ▶ Other ways to encode qualitative predictors produce the same fit \hat{f}_n , but the coefficients have different interpretations.
- ▶ Hypothesis tests derived from these dummy indicator are affected by the choice.
 - ▶ **Solution:** To check whether ethnicity is important, use an F -test for the hypothesis $\beta_{Asian} = \beta_{Caucasian} = 0$.



So far, we have:

- ▶ Defined Multiple Linear Regression
- ▶ Discussed how to estimate model parameters
- ▶ Discussed how to test the importance of variables
- ▶ Described one approach to choose a subset of variables
- ▶ Explained how to code dummy indicators

What are some potential issues?



- ▶ Interactions between predictors
- ▶ Non-linear relationships
- ▶ Correlation of error terms
- ▶ Non-constant variance of error (heteroskedasticity)
- ▶ Outliers
- ▶ High leverage points
- ▶ Collinearity
- ▶ Mis-specification

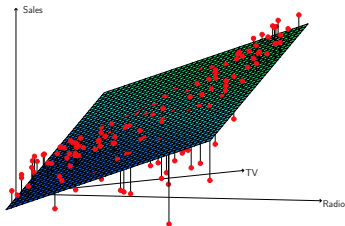


Linear regression has an *additive* assumption, e.g.:

$$\text{sales} = \beta_0 + \beta_1 \cdot \text{tv} + \beta_2 \cdot \text{radio} + \epsilon \quad (36)$$

e.g. An increase of \$ 100 dollars in TV ads correlates to a fixed increase in sales, independent of how much you spend on radio ads.

If we visualize the residuals, it is clear that this is false:





One way to deal with this:

- ▶ Include multiplicative variables (aka interaction variables) in the model

$$sales = \beta_0 + \beta_1 \cdot tv + \beta_2 \cdot radio + \beta_3 \cdot (tv \times radio) + \epsilon \quad (37)$$

- ▶ Makes the effect of TV ads dependent on the radio ads (and vice versa)
- ▶ The *interaction variable* is high when both tv and radio are high



Two ways of including interaction variables (in R):

- ▶ Create a new variable that is the product of the two
- ▶ Specify the interaction in the model formula

```
> lm.fit=lm(Sales~.+Income:Advertising+Price:Age,data=Carseats)
> summary(lm.fit)

Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data =
  Carseats)

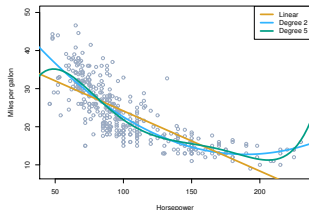
Residuals:
    Min       1Q   Median       3Q      Max
-2.921  -0.750   0.018   0.675   3.341

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.575565    1.008747   6.52  2.2e-10 ***
CompPrice    0.092937    0.004118  22.57  < 2e-16 ***
Income       0.010894    0.002604   4.18  3.6e-05 ***
Advertising  0.070246    0.022609   3.11  0.00203 **
Population   0.000159    0.000368   0.43  0.66533
Price       -0.100806    0.007440 -13.55  < 2e-16 ***
ShelveLocGood  4.848676    0.152838  31.72  < 2e-16 ***
ShelveLocMedium 1.953262    0.125768  15.53  < 2e-16 ***
Age          -0.057947    0.015951  -3.63  0.00032 ***
Education    -0.020852    0.019613  -1.06  0.28836
UrbanYes     0.140160    0.112402   1.25  0.21317
USYes       -0.157557    0.148923  -1.06  0.29073
Income:Advertising 0.000751    0.000278   2.70  0.00729 **
Price:Age     0.000107    0.000133   0.80  0.42381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Scatterplots between X and Y may reveal non-linear relationships

- **Solution:** Include polynomial terms in the model



$$\begin{aligned} \text{MPG} = & \beta_0 + \beta_1 \cdot \text{horsepower} \\ & + \beta_2 \cdot \text{horsepower}^2 \\ & + \beta_3 \cdot \text{horsepower}^3 + \dots + \epsilon \end{aligned} \quad (38)$$

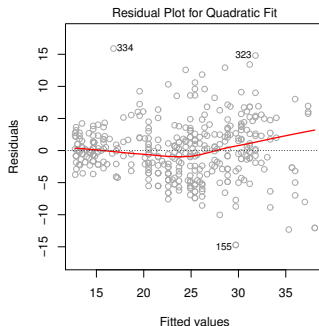
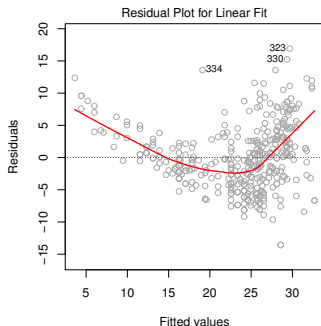


In 2 or 3 dimensions, this is easy to visualize. What do we do when we have too many predictors?



In 2 or 3 dimensions, this is easy to visualize. What do we do when we have too many predictors?

Plot the residuals against the response and look for a pattern:





We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \epsilon_i : \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad (39)$$



We assumed that the errors for each sample are independent:

$$y_i = f(x_i) + \epsilon_i : \epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2) \quad (39)$$

When it doesn't hold:

- Invalidates any assertions about Standard Errors, confidence intervals, and hypothesis tests

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

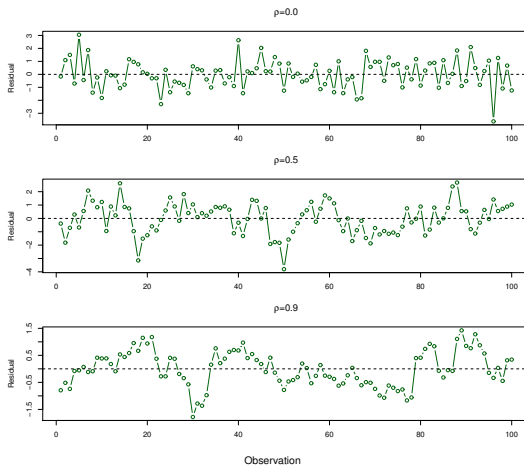


Examples of when this happens:

- ▶ *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.
- ▶ *Clustered data*: Study on 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.



Simulations of time series with increasing correlations on ϵ_i .

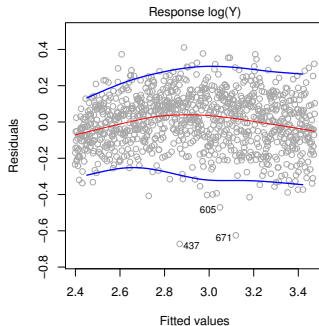
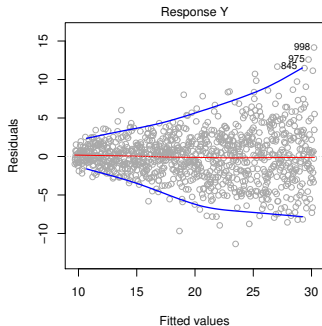


Non-constant variance of error (heteroskedasticity)



The variance of the error depends on the input value.

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

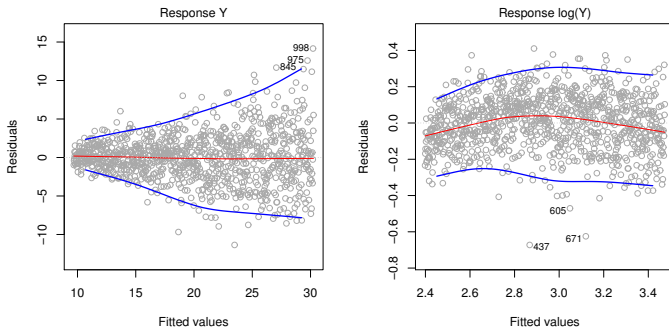


Non-constant variance of error (heteroskedasticity)



The variance of the error depends on the input value.

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.



[1] ISL. Chapters 3.

[2] ESL. Chapters 3.