Linear Regression

Fittin' lines

galvanize

Objectives

- Review Linear Regression
- Explain the difference between residuals and irreducible error
- Name some regression diagnostics
- Name model test statistics and when to use which
- State how we can deal with nonlinearity

Linear Regression

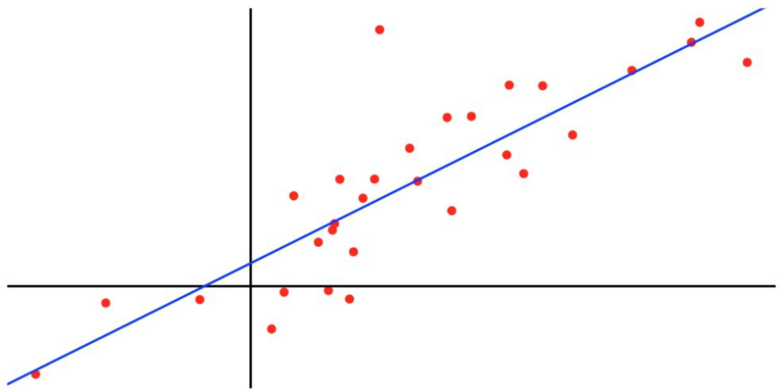
- Goal: predict a continuous output variable (Y) from a set of predictor variables (X)
- *Parametric* model (vs. *non-parametric* models)
- Simple and interpretable
- Trying Linear Regression before trying more complicated models is often a good idea
- Example: predict house price based on # sqft and neighborhood

Simple Linear Regression

galvanize

With linear regression we "train" a model on some data.

Sometimes called learning, estimation, model fitting.



X	Y
Stock Quote	Future Stock Price
% of Diabetes	Mortality Rate
Historic Web Logs	Page Views
Airplane Flight Status	Arrival Time
Anything!	Anything!

X data are generally called "features."

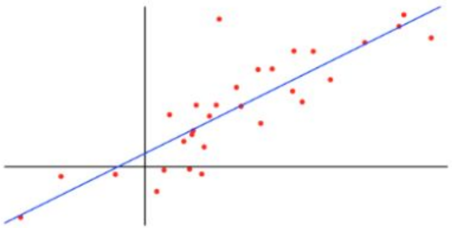
Sometimes called covariates, independent variables, inputs.

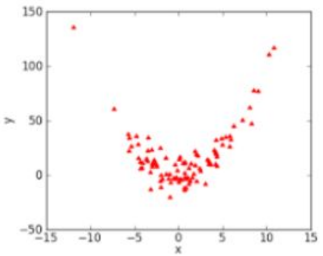
Y data are generally called "targets."

Sometimes called labels, dependent variable, outputs.

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$





- $Y = \beta_0 + \beta_1 X + \varepsilon = f(X) + \varepsilon$
- $f(X) = \beta_0 + \beta_1 X$ is the true underlying dependency of Y on X
- ε is the irreducible error coming from factors that we have not or cannot measure
- $\hat{f}(X) = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is out model's estimate for the true f(X)
- We try to minimize f(X) - f^(X), this is called the reducible error
- We can never get rid of ε, unless we find new features that are correlated to them

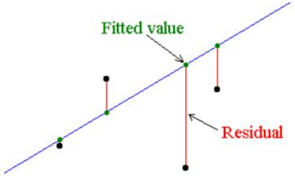
Simple Linear Regression

- $Y = \beta_0 + \beta_1 X + \varepsilon = f(X) + \varepsilon$
- $\hat{y} = \hat{f}(X = x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is your model's estimate for datapoint X=x
- $e_i = y_i - \hat{y}_i$ is your model's error for datapoint x_i
- e_i contains the reducible and the irreducible error
- For the perfect model $e_i = \varepsilon_i$
- For linear regression we often use the residual sum of squares to assess the quality of the fit
- $RSS = e_1^2 + e_2^2 + \dots + e_n^2$

Simple Linear Regression

Linear Regression is often called **Ordinary Least Squares (OLS) Regression** because the model simply finds coefficients that **minimize the sum total squared distance (residuals)** between each data point and the line.

$$e_i = y_i - \hat{y}_i$$



Want these to be small

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$

Typically square them!
(though absolute value is an alternative)

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Multiple Linear Regression (fitting hyperplanes)

With Multiple Linear Regression we can combine many features into a single model.

Model

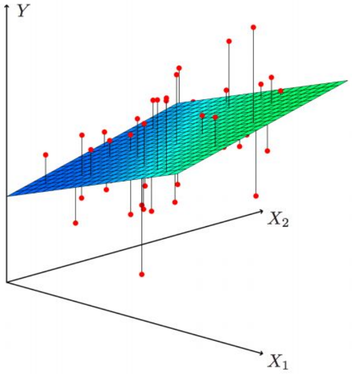
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

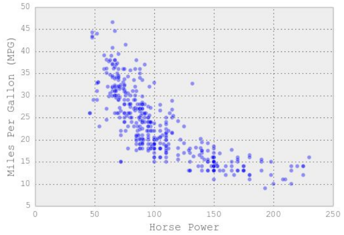
Fitted Value

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Residual Sum of Squares

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$





We can also have non-linear features, like X and X^2.

Ordinary Least Squares (OLS)

Minimize the RSS in terms of matrix algebra:

$$\mathbf{Y}_{n\times 1} \quad = \quad \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

X is called the “design” matrix, **β** is the parameter vector and **y** the target vector

Ordinary Least Squares (OLS)

Minimize the RSS in terms of matrix algebra: $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$

Problem: find **β** such that $S(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|^2$ is minimized

Some matrix calculus yields: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$

For 1D linear regression this yields

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

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

$\hat{\beta}_1$ simply the covariance of x and y (normalized by variance of x)

Model Accuracy

Residual Sum of Squares

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

Not great...

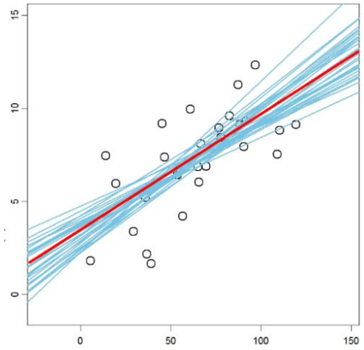
R-Squared, or "Proportion of Variance Explained"

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

where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

☺ Nice interpretation
Independent of scale of y

Test statistics: t-test (Is this variable important?)



$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

	Recall	**Here**
Setup Hypothesis	H_0: $\mu = 100$	$H_0 : \beta_1 = 0$
Sample Statistic	$\bar{x}$	$\hat{\beta}_1$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
Confidence Interval	$(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}})$	$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$

Test if X has effect on Y

Test statistics: F-test (Is ANY variable important?)

F-test compares model with just a subset of m predictors to full model with p predictors (m<p)

Ex: predict MPG (Y) from set of variables:

Full model: $Y = \beta_0 + \beta_{weight} + \beta_{height} + \beta_{model} + \beta_{year} + \beta_{color}$

Reduced model: $Y = \beta_0 + \beta_{weight} + \beta_{year} + \beta_{color}$

Calculate F-statistic (ratio of variance left unexplained by reduced model vs full model)

$$F = \frac{(RSS_{reduced} - RSS_{full})/(p_{full} - p_{reduced})}{RSS_{full}/(n - p_{full} - 1)}$$

where F has degrees of freedom (p_full - p_reduced), (n - p_full – 1)

Test statistics: F-test (Is ANY variable important?)

$$F = \frac{(RSS_{reduced} - RSS_{full})/(p_{full} - p_{reduced})}{RSS_{full}/(n - p_{full} - 1)}$$

where F has degrees of freedom (p_full - p_reduced), (n - p_full – 1)

If F is large, the dropped parameters are important

If you drop just one parameter and evaluate the p-value from the F-table you get back the p-value for the t-test for that parameter!

The statsmodels F-statistic and p-value correspond to dropping all parameters (null model vs full model), it will tell you if at least one of the parameters of the full set is important

Model Interpretation (statsmodels)

galvanize

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
Omnibus:                        0.655   Durbin-Watson:                   2.896
Prob(Omnibus):                  0.721   Jarque-Bera (JB):                0.360
Skew:                           0.207   Prob(JB):                        0.835
Kurtosis:                       3.026   Cond. No.                         221.
==============================================================================
```

Model Interpretation (statsmodels)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
Omnibus:                        0.655   Durbin-Watson:                   2.896
Prob(Omnibus):                  0.721   Jarque-Bera (JB):                0.360
Skew:                           0.207   Prob(JB):                        0.835
Kurtosis:                       3.026   Cond. No.                         221.
==============================================================================
```

Proportion of Variance Explained by model is 93.3%

Measure of the significance of the fit …my model isn't utterly useless ☺

There is an approximately 95% chance that [0.275, 0.693] will contain the true value of β_2

Each coefficient is really significant. Can also think of this as a Partial F-test.

"The average effect on Y of a one unit increase in X_2, holding all other predictors (X_1 & X_3) fixed, is 0.4836"

- However, interpretations are generally pretty hazardous due to correlations among predictors.
- p-values for each coefficient ≈ 0, so might be okay here

Note: Magnitude of the Beta coefficients is NOT how to determine whether predictor contributes. Why?

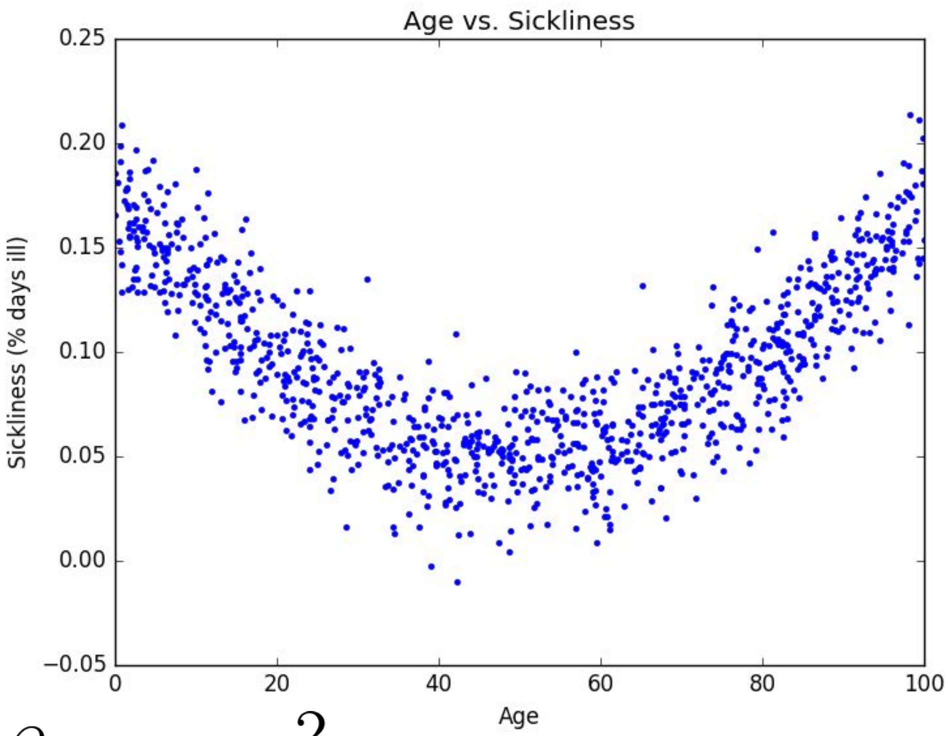
Assumptions

- Linear relationship!
- Constant variance (homoscedasticity)
- Independence of errors
- Normality of errors
- Lack of multicollinearity

Non-linearity

We can make linear regression non-linear by inserting extra "interaction" features or higher-order features.

As you add more features, R^2 will only go up.



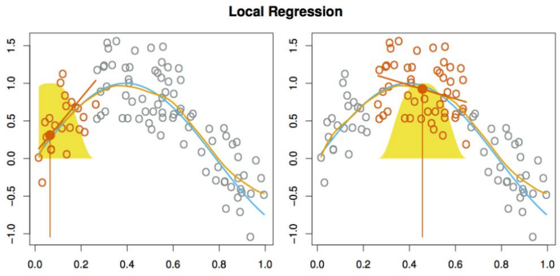
Example:

$$Y = \beta_0 + \beta_1 * \text{age}$$

$$Y = \beta_0 + \beta_1 * \text{age} + \beta_2 * \text{age}^2$$

Many other methods for non-linear regression exist but will not be discussed

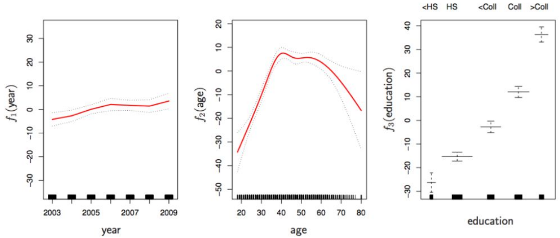
ISLR and ESLR go into them in more detail: GAMs, local regression, splines etc



Local Regression

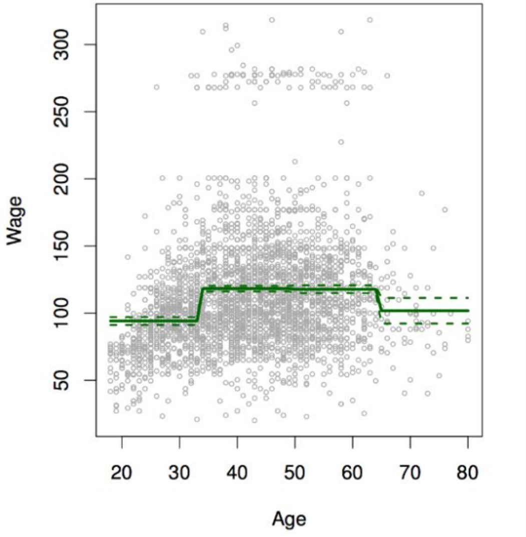
- Use sliding weight function, make separate linear fits over range of X

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i.$$



Generalized Additive Models

- Just add up contributing effects



Step Functions

Multicollinearity

Multicollinearity occurs when two or more X features are correlated to each other.

For example $x_2 = 2*x_1$

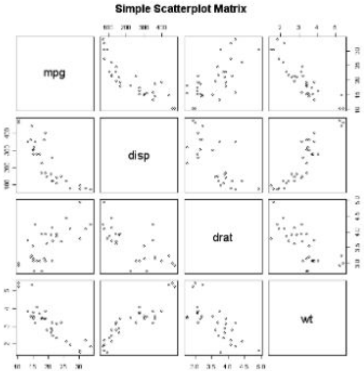
What happens	What you do
• The uncertainty in the model coefficients becomes large. • Does not affect the model accuracy, only the interpretability of the coefficients.	• Use correlation matrix to look for pairwise correlations. • Use VIF for more complicated relationships. • Remove (but make note of) any predictor that is easily determined by the remaining predictors.

Multicollinearity diagnosis

galvanize

- Identification by correlation matrix and pairwise scatter plots

	DJIA	S&P 500	Nasdaq	Canada	Mexico	Brazil	Stoxx 50	FTSE 100	CAC 40	DAX	IBEX	Italy	Netherlands	Sweden	Switzerland	Nikkei	Hang Seng	Australia
DJIA		0.97	0.85	0.57	0.56	0.52	0.52	0.48	0.51	0.56	0.49	0.50	0.50	0.42	0.42	0.09	0.11	0.07
S&P 500	0.97		0.91	0.62	0.58	0.55	0.50	0.47	0.50	0.55	0.48	0.50	0.49	0.41	0.41	0.09	0.11	0.05
Nasdaq	0.85	0.91		0.58	0.56	0.52	0.48	0.43	0.48	0.54	0.47	0.48	0.48	0.42	0.38	0.14	0.16	0.07
Canada	0.57	0.62	0.58		0.53	0.53	0.42	0.45	0.41	0.41	0.42	0.42	0.39	0.37	0.35	0.17	0.22	0.17
Mexico	0.56	0.58	0.56	0.53		0.56	0.42	0.42	0.44	0.43	0.43	0.44	0.39	0.38	0.38	0.17	0.25	0.17
Brazil	0.52	0.55	0.52	0.53	0.56		0.33	0.35	0.32	0.34	0.34	0.34	0.29	0.30	0.28	0.17	0.22	0.15
Stoxx 50	0.52	0.50	0.48	0.42	0.42	0.33		0.92	0.94	0.89	0.87	0.88	0.92	0.78	0.86	0.26	0.30	0.24
FTSE 100	0.48	0.47	0.43	0.45	0.42	0.35	0.92		0.86	0.80	0.80	0.82	0.84	0.73	0.78	0.26	0.30	0.26
CAC 40	0.51	0.50	0.48	0.41	0.44	0.32	0.94	0.86		0.89	0.88	0.89	0.92	0.78	0.84	0.28	0.32	0.25
DAX	0.56	0.55	0.54	0.41	0.43	0.34	0.89	0.80	0.89		0.83	0.84	0.86	0.75	0.77	0.26	0.29	0.21
IBEX	0.49	0.48	0.47	0.42	0.43	0.34	0.87	0.80	0.88	0.83		0.84	0.83	0.75	0.77	0.27	0.32	0.26
Italy	0.50	0.50	0.48	0.42	0.44	0.34	0.88	0.82	0.89	0.84	0.84		0.85	0.74	0.78	0.24	0.29	0.23
Netherlands	0.50	0.49	0.48	0.39	0.39	0.29	0.92	0.84	0.92	0.86	0.83	0.85		0.75	0.82	0.27	0.30	0.23
Sweden	0.42	0.41	0.42	0.37	0.38	0.30	0.78	0.73	0.78	0.75	0.75	0.74	0.75		0.75	0.29	0.33	0.27
Switzerland	0.42	0.41	0.38	0.35	0.38	0.28	0.86	0.78	0.84	0.77	0.77	0.78	0.82	0.75		0.29	0.32	0.29
Nikkei	0.09	0.09	0.14	0.17	0.17	0.17	0.26	0.26	0.28	0.26	0.27	0.24	0.27	0.29	0.29		0.52	0.49
Hang Seng	0.11	0.11	0.16	0.22	0.25	0.22	0.30	0.30	0.32	0.29	0.32	0.29	0.30	0.33	0.32	0.52		0.48
Australia	0.07	0.05	0.07	0.17	0.17	0.15	0.24	0.26	0.25	0.21	0.26	0.23	0.23	0.27	0.29	0.49	0.48	



Downside is can only pick up pairwise effects ☹

- Variance inflation factor (VIF) calculation:
  - Run Linear regression for each predictor as target as function of all other predictors (p times)

$$X_1 = \alpha_2 X_2 + \alpha_3 X_3 + \cdots + \alpha_k X_k + c_0 + e$$

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

  - Rule of thumb: collinearity is high if $\mathrm{VIF}(\hat{\beta}_i) > 10$

Inspecting residuals

$$e_i = y_i - \hat{y}_i$$

- They will be more useful if we standardize them:
- $r_i = e_i/\sigma$, with σ^2 the true population variance, which is unknown but can be estimated by the mean squared error (MSE)
- e_i/sqrt(MSE) is called the semi-studentized residual
- A better estimate of the real variance is $\widehat{V}(e_i) = MSE(1 - h_{ii})$
- Which gives us the studentized residual: $r_i = \dfrac{e_i}{\sqrt{MSE(1 - h_{ii})}}$

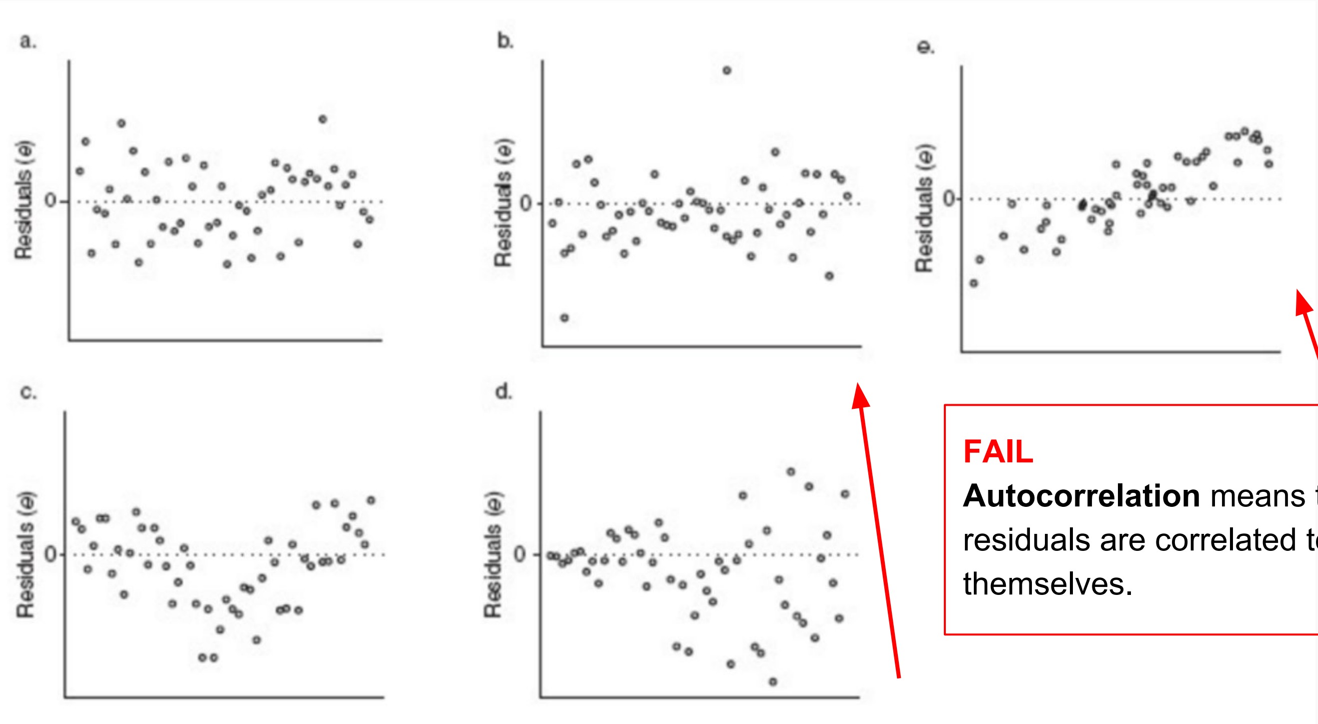
h_{ii} is called the "self-influence" and are diagonal elements of the "hat matrix" H (also "projection matrix")

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$$

$$e = y - \hat{y} = y - Hy = (I - H)y$$

Residual Plots Allow Us to Check Assumptions

galvanize



PASS
Residuals are **normally distributed.**

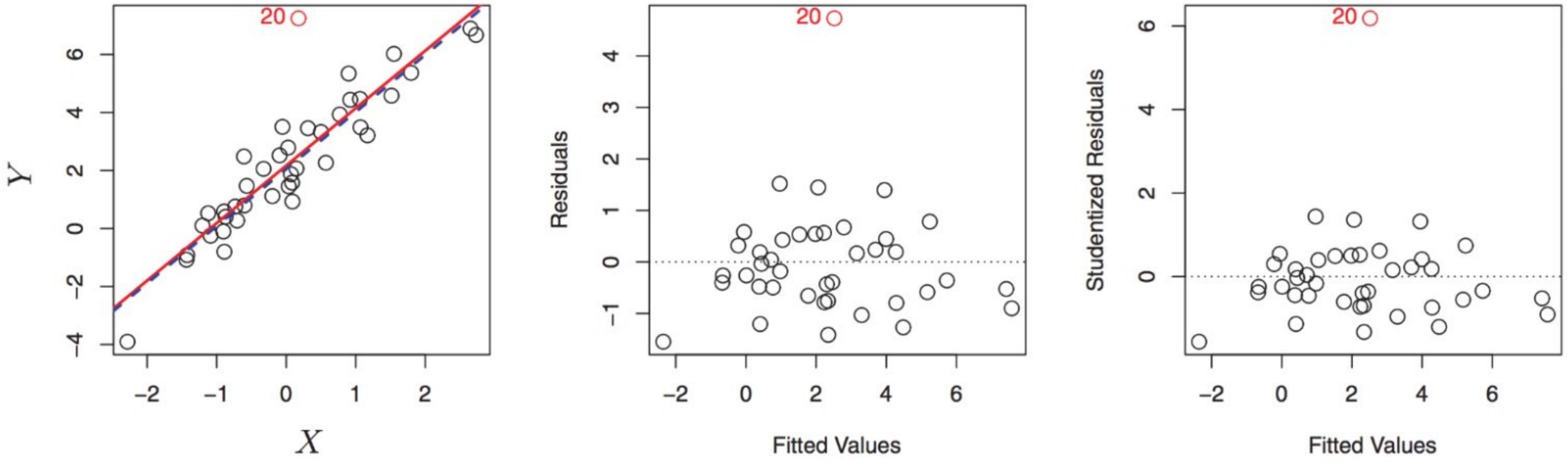
FAIL
Non-Linear relationship between X and y.

FAIL
Autocorrelation means the residuals are correlated to themselves.

FAIL
Heteroscedasticity means the variance is not constant relative to X.

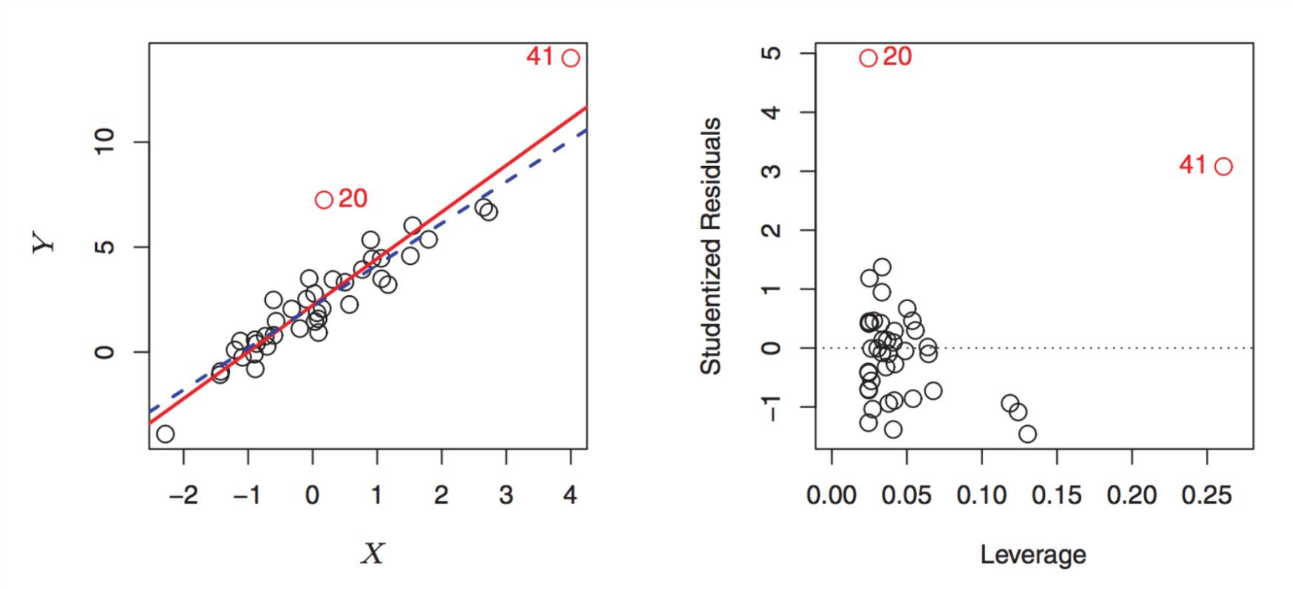
Outliers

- Outliers are data points that are far from their predicted values
- Can be identified by inspecting residuals
- If studentized residual >> 2 or << -2, we consider them outliers



High leverage points

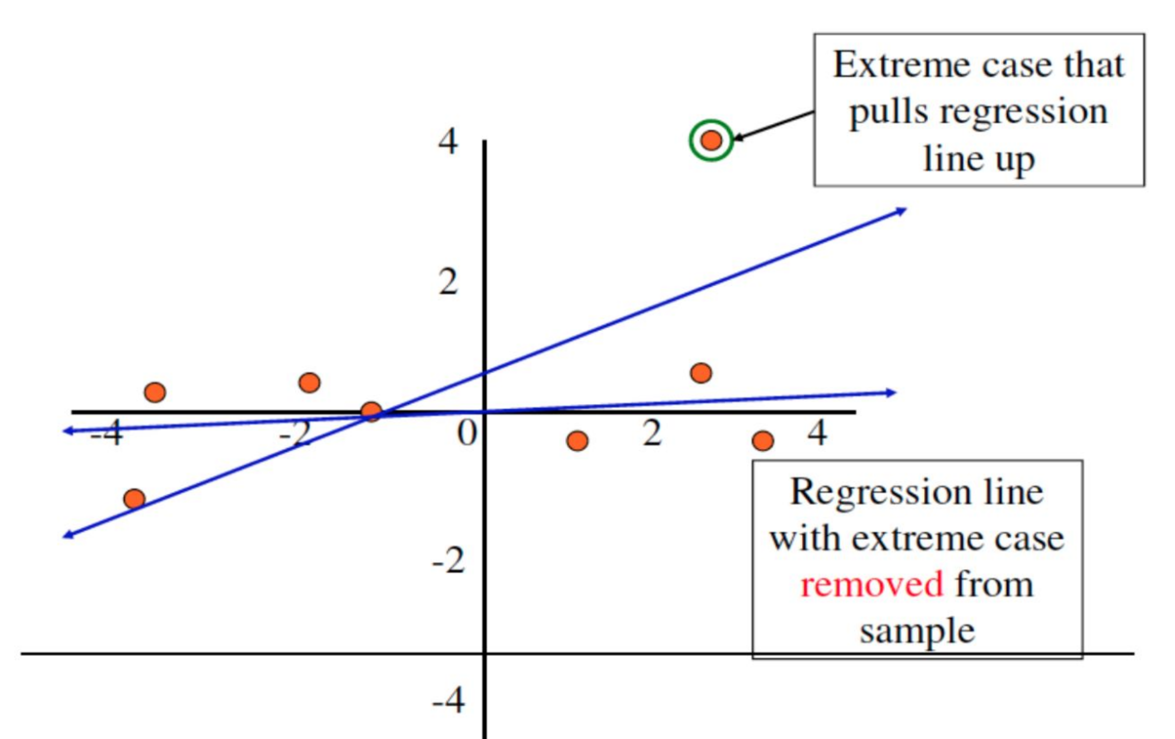
- Outliers are different from high leverage points
- A high leverage point has an extreme X value (far from the rest of the data)
- Commonly measured with the diagonal elements of the hat matrix H
- In the following dataset, 20 is an outlier, 41 is a leverage point AND outlier



$$H = X(X^TX)^{-1}X^T$$

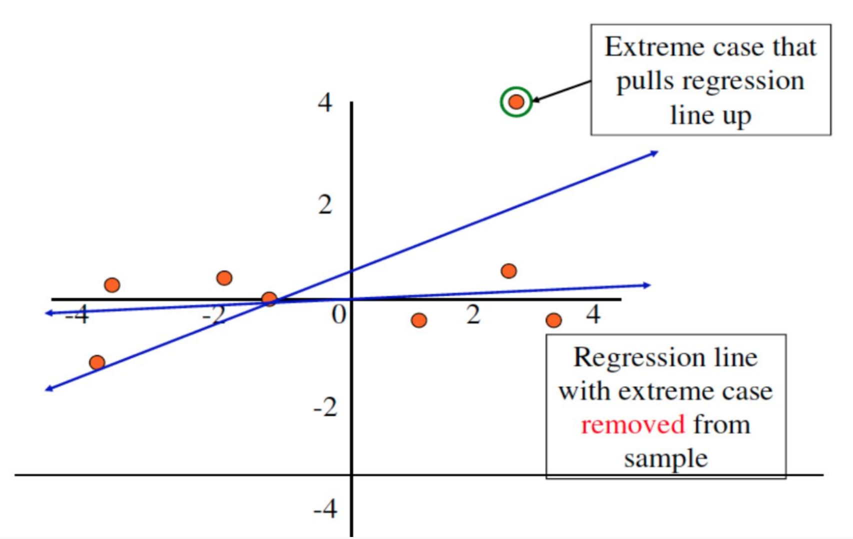
Influential points

- Observations that are outliers and have high leverage tend to be influential
- Their removal from the data greatly affects the slope of the regression line



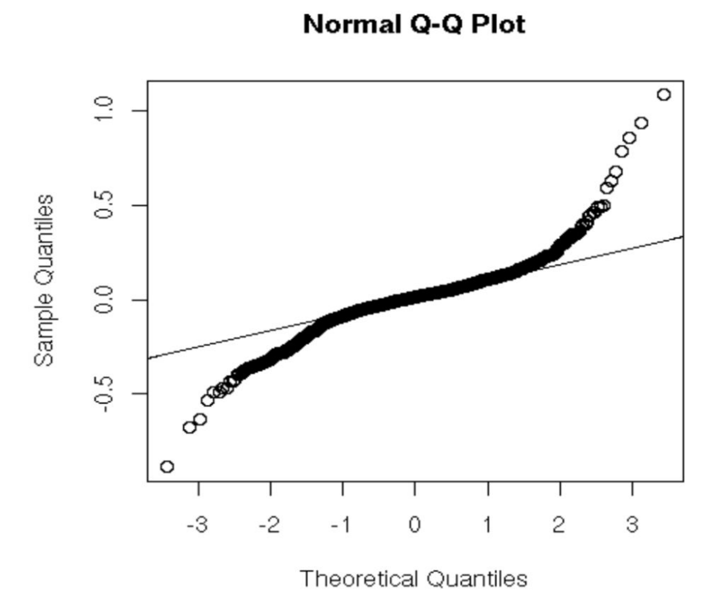
Influential points

- An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.
- Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.



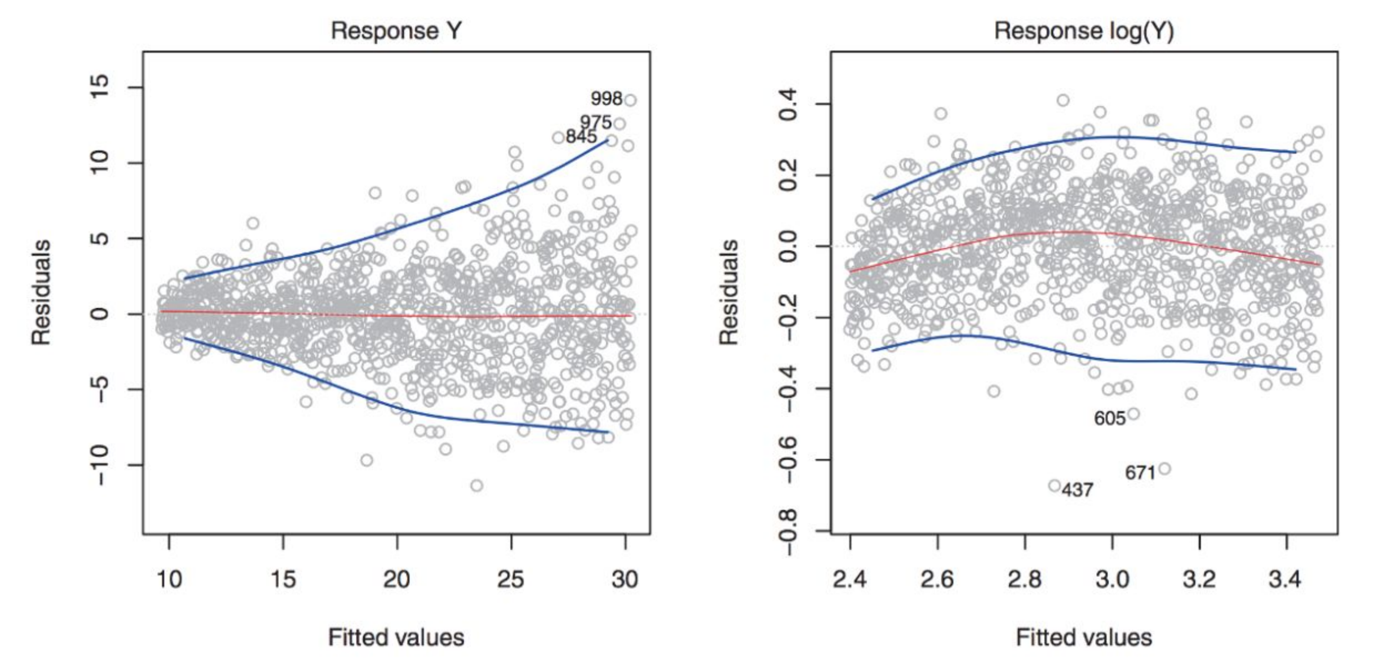
Normality of errors

- Normality assumption allows us to do hypothesis testing (t-tests) on our parameters, and construct confidence intervals
- Ways to check: QQ-plots of residuals against normal distribution
- Ways to fix: transformation of Y (e.g. log(Y))



Heteroscedasticity of errors

- Variance changes depending on X
- Fix: transform Y (log(Y) or sqrt(Y) for example)



Linear Regression

Fittin' lines

galvanize

Objectives

- Review Linear Regression
- Explain the difference between residuals and irreducible error
- Name some regression diagnostics
- Name model test statistics and when to use which
- State how we can deal with nonlinearity

Afternoon

Categorical variables and interactions

galvanize

Objectives

- State how to deal with categorical variables
- State how to include interactions into your model
- State why model validation is important and how it works

Categorical variables

- Independent variable might not be numerical
- Ex: using "gender" and "ethnicity" to predict credit card balances
- Solution: Create "dummy variable" that takes on value of 0 or 1

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

$$y_i = \beta_0 + \beta_1 \underline{x_i} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Categorical variables with more than 2 levels

- Ethnicity has 3 levels: African-American, Asian, Caucasian

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is } \underline{\text{Asian}} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is } \underline{\text{Caucasian}} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 \underline{x_{i1}} + \beta_2 \underline{x_{i2}} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

Categorical variables with more than 2 levels

$$y_i = \beta_0 + \beta_1 \underline{x_{i1}} + \beta_2 \underline{x_{i2}} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

Data

Ones	Ethnicity
1	AA
1	Asian
1	Asian
1	Caucasian
1	AA
1	AA
1	Asian
1	Caucasian
1	AA
…	…

→

Recode Design Matrix

Ones	Asian	Caucasian
1	0	0
1	1	0
1	1	0
1	0	1
1	0	0
1	0	0
1	1	0
1	0	1
1	0	0
…	…	…

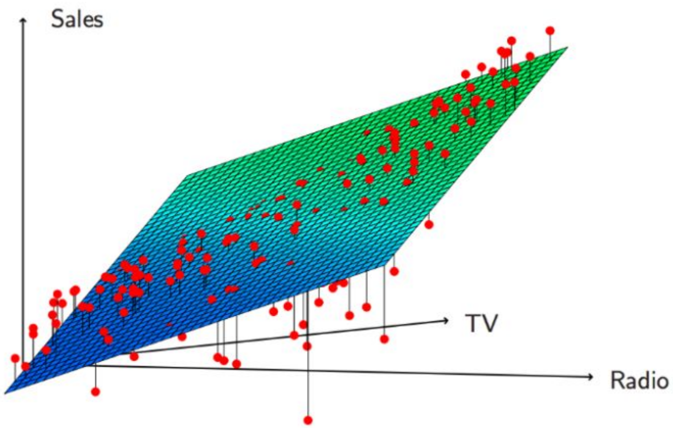
Categorical variables with more than 2 levels

$$y_i = \beta_0 + \beta_1 \underline{x_{i1}} + \beta_2 \underline{x_{i2}} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is AA.} \end{cases}$$

- β_0 does not represent the general baseline anymore, but the baseline for group 0 (AA)
- β_1 represents the difference of the baseline of group 1 (Asian) to group 0 (AA)
- β_2 represents difference group 0 and group 2
- What does it mean if β_1 = -20.3 ?
- What do we do if we want to use Caucasians as the baseline?

Interactions/Synergy

$$\widehat{\texttt{sales}} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper}$$



Suggests synergy between TV and Radio

- Synergy between radio and TV means that spending 50K on both radio and TV is better for sales than spending 100K on only one of them
- How can we model this?

Interactions of quantitative variables

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \underline{\beta_3 \times (\texttt{radio} \times \texttt{TV})} + \epsilon \\ &= \beta_0 + (\underline{\beta_1 + \beta_3 \times \texttt{radio}}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon.\end{aligned}$$

Results:

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	$\boxed{< 0.0001}$ ← Improvement!

- Changing radio will change the slope of TV!

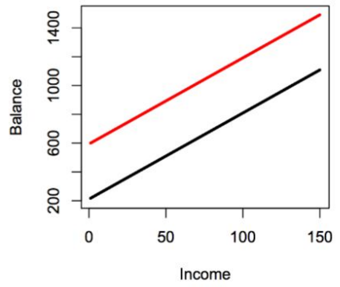
Interactions of categorical variables

- Interaction of student (categorical) and income (quantitative)

No Interaction $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i$

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

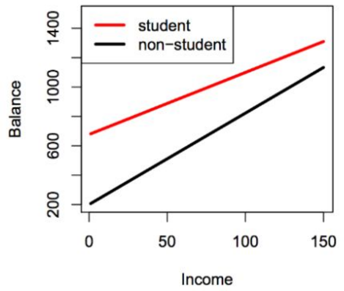
$$= \underline{\beta_1} \times \texttt{income}_i + \begin{cases} \underline{\beta_0 + \beta_2} & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.} \end{cases}$$



With Interaction $balance_i = \beta_0 + \beta_1 * income_i + \beta_2 * student_i + \beta_3 * income_i * student_i$

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases}$$

$$= \begin{cases} \underline{(\beta_0 + \beta_2)} + \underline{(\beta_1 + \beta_3)} \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}$$



Is R^2 all that matters?

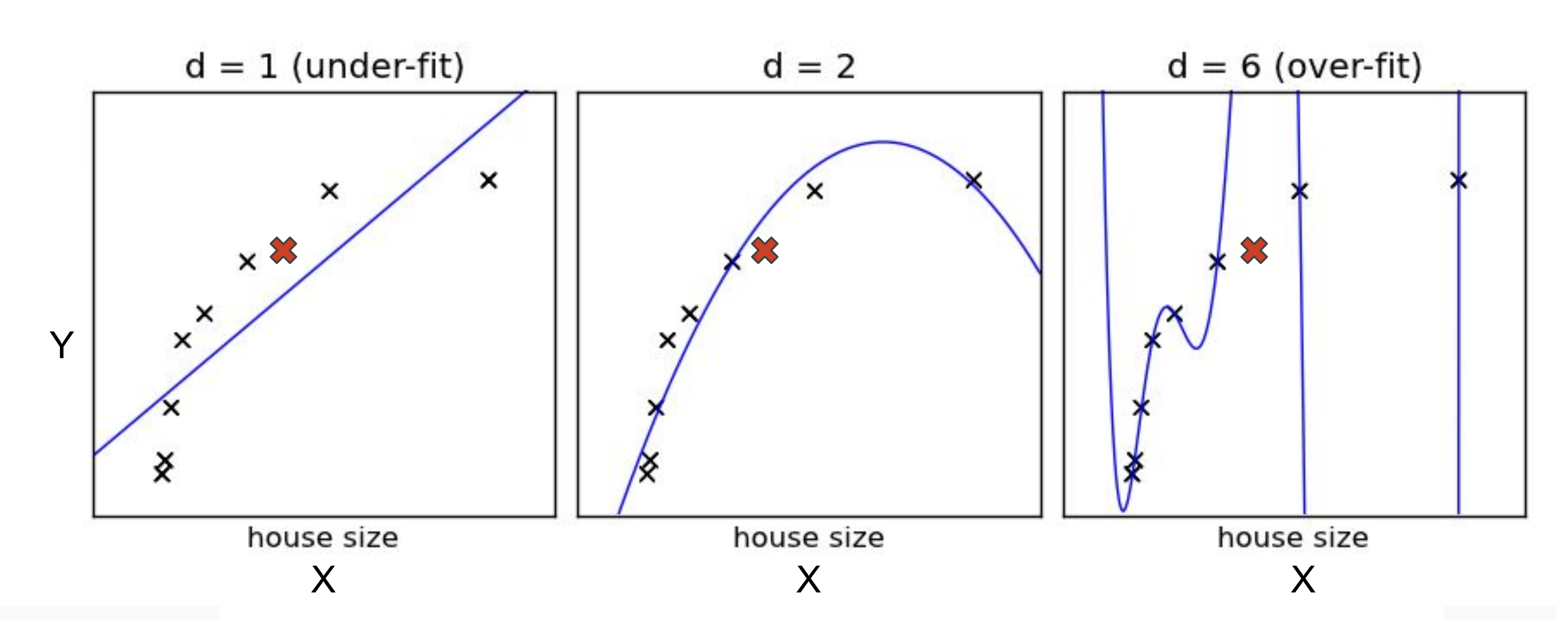
We *could* just keep inserting interaction features until $R^2 = 1$.

Boom. I solved data science. Here's my idea:

```
def train_super_awesome_perfect_model (X, y):
    while True:
        model = LinearRegression()
        model.fit(X, y)
        if calculate_r2(model, X, y) >= 0.999:
            return model
        else:
            X = insert_random_interaction_feature(X)
```

Why is this a bad idea?

Oh the woes of overfitting...



What's bad about the first model?

What's bad about the second model?

What's bad about the third model?

Model Validation

Main idea: **Don’t use all your data for training.**

Instead: **Split your data into a “training set” and a “validation set”.**

randomize!
(why?)

training set

where to split?

validation set

Validation techniques will be covered tomorrow!

Questions

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

- Name and explain all the elements of the above equation and its meaning
- How do we assess the quality of a model/fit
    - How are the different measures related?
- How do you compare a submodel of a model to the model?
    - How does this relate to the t-statistic for individual parameter?
    - How does this relate to the standard F-statistic and p-value printed by statsmodels?
- How do we account for categorical variables?
    - How do we change the baseline?
- What is an interaction/synergy?
    - How do we account for it?

Afternoon

Categorical variables and interactions

galvanize

Objectives

- State how to deal with categorical variables
- State how to include interactions into your model
- State why model validation is important and how it works