ExamPA.net

Generalized Linear Models Part II

Lesson

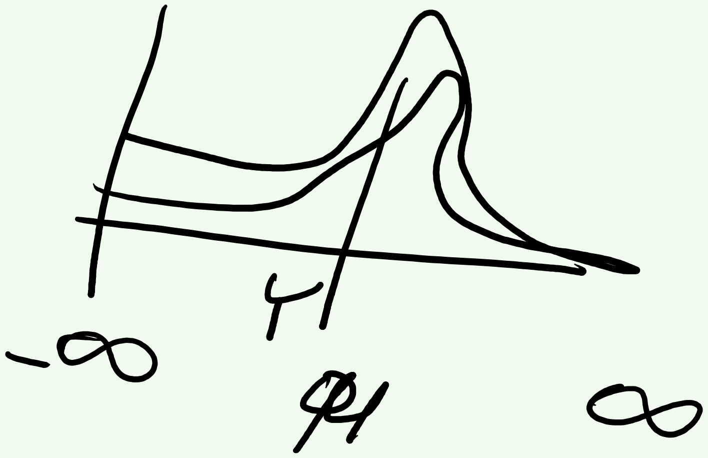
Learning Objectives

6. **Topic: Generalized Linear Models**
Learning Objectives
The Candidate will be able to describe and select a Generalized Linear Model (GLM) for a given data set and regression or classification problem.
Learning Outcomes
The Candidate will be able to: a) Implement ordinary least squares regression in R and understand model assumptions. b) Understand the specifications of the GLM and the model assumptions. c) Create new features appropriate for GLMs. d) Interpret model coefficients, interaction terms, offsets, and weights. e) Select and validate a GLM appropriately. f) Explain the concepts of bias, variance, model complexity, and the bias-variance trade-off. g) Select appropriate hyperparameters for regularized regression.

Assumptions of OLS

Random component

#1 $Y|X \sim N(\mu = E[Y|X])$

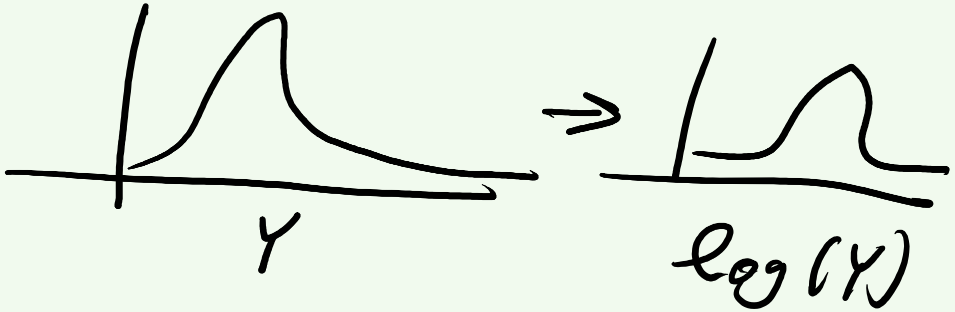


#2

systemic component

$$\mu = \mu(X) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

$$Y|X \sim N(X\beta, \sigma^2 I)$$



Assumptions of GLMs

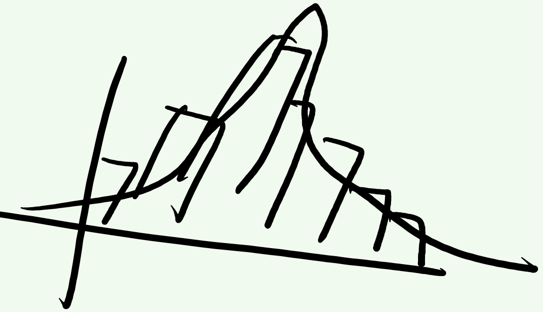
#1 $Y/X \sim \underbrace{\text{exponential}}_{\text{Bin, Gaussian, Gamma, etc}}$

#2 $g(\underbrace{\mu(x)}_{\text{Mean of } Y}) = X\beta$

$Y \sim \Gamma(\alpha, \beta)$

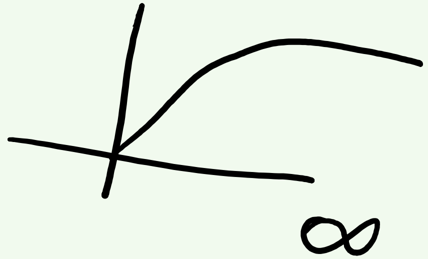
$\mu_i = \alpha_i \beta_i$

	\$	
Sam	\$5,000	$\mu = \$4,500$
Bob	\$30,000	$\mu = \$25,000$



Link Functions for Regression

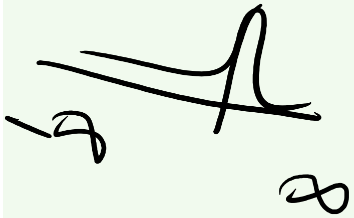
Name	Link Function	Mean	Range of Mean
Identity	$XB = z = \mu$	$\mu = z$	$(-\infty, +\infty)$
Log	$z = \log(\mu)$	$\mu = \exp(z)$	$(0, +\infty)$
Inverse	$z = 1/\mu$	$\mu = \frac{1}{z}$	$(-\infty, +\infty)$
Inverse Squared	$z = 1/\mu^2$	$\mu = \frac{1}{\sqrt{z}}$	$(0, +\infty)$
Square root	$z = \sqrt{\mu}$	$\mu = z^2$	$(0, +\infty)$



Response Families for Regression

$g(\mu) = XB$ $g^{-1}(XB) = \mu$

Distribution	Range	Skewed
Gaussian	$(-\infty, +\infty)$	*No*
Binomial	$\{0,1\}$	*NA*
Gamma	$(0, +\infty)$	*Yes*
Inverse Gaussian	$(0, +\infty)$	*Yes*
Poisson	$\{0,1,2,3,4,5\}$	*Yes*

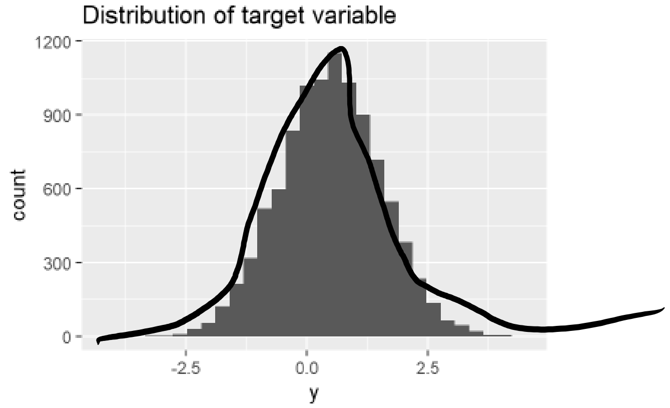


Example: Gaussian/Identity

Q1 - Choose a distribution and link function.

Determine the best distribution and link function to use. Justify your choice of distribution based on the business problem and data and use only that combination for all further work. Test several combinations and select the one with the best AIC, QQ-plot, and graphs of residuals vs. fitted.

The distribution of the target variable Y is shown below, for $n = 10,000$ observations.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9828 -0.2085  0.4971  0.4904  1.1956  4.4772
```

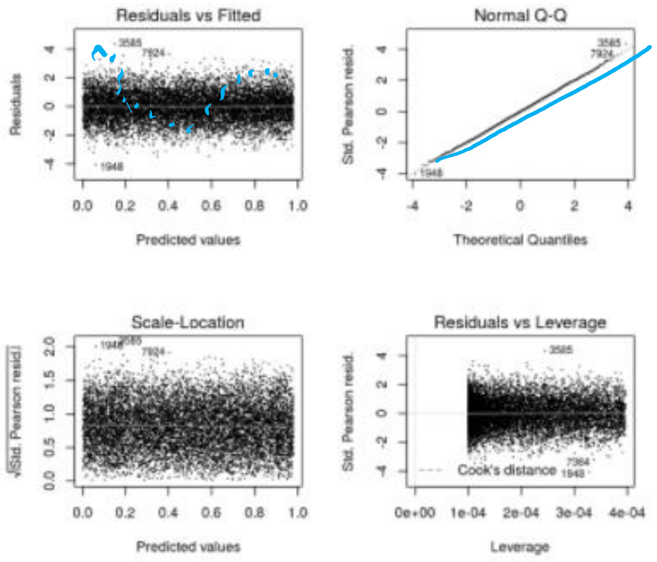
Sandbox | Start Over | Hints | Run Code

```
glm <- glm(formula=y ~ x, family = gaussian(link = "identity"), data = glmdata1)
summary(glm)
par(mfrow = c(2,2))
plot(glm, cex= 0.1)
#lakjsdfasd
```

Output

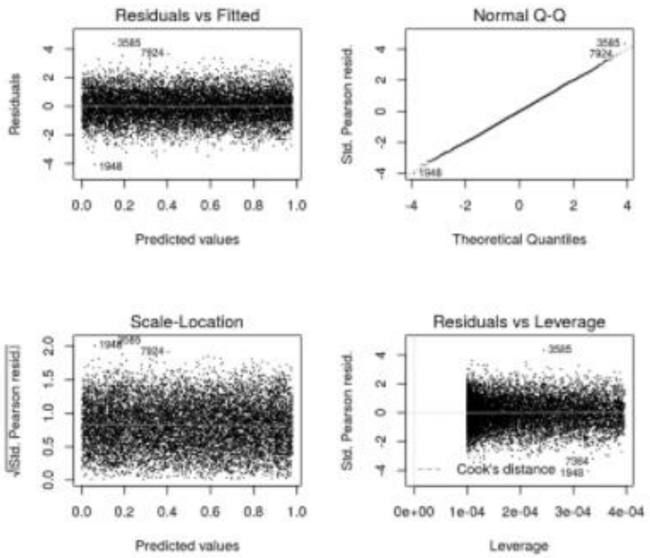
```
Call:
glm(formula = y ~ x, family = gaussian(link = "identity"), data = glmdata1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0451  -0.6766   0.0014   0.6781   4.3284

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00324    0.01996   0.162    0.871
x            0.97643    0.03456  28.250   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.00968)

    Null deviance: 10901  on 9999  degrees of freedom
Residual deviance: 10095  on 9998  degrees of freedom
AIC: 28479

Number of Fisher Scoring iterations: 2
```

$\hat{y} = 0.003 + 0.97x$



Output

```
Call:
glm(formula = y ~ x, family = gaussian(link = "identity"), data = glmdata1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0451  -0.6766   0.0014   0.6781   4.3284

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00324    0.01996   0.162    0.871
x            0.97643    0.03456  28.250   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.00968)

    Null deviance: 10901  on 9999  degrees of freedom
Residual deviance: 10095  on 9998  degrees of freedom
AIC: 28479

Number of Fisher Scoring iterations: 2
```

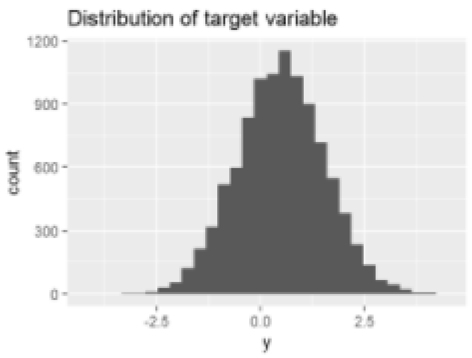


Example: Gaussian/log?

Q1 - Choose a distribution and link function.

Determine the best distribution and link function to use. Justify your choice of distribution based on the business problem and data and use only that combination for all further work. Test several combinations and select the one with the best AIC, QQ-plot, and graphs of residuals vs. fitted.

The distribution of the target variable Y is shown below, for $n = 10,000$ observations.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9828 -0.2085  0.4971  0.4904  1.1956  4.4772
```

Sandbox | Start Over | Hints | Run Code

```
glm <- glm(formula=y ~ x, family = gaussian(link = "log"), data = glmdata1)
summary(glm)
par(mfrow = c(2,2))
plot(glm, cex= 0.1)
#lakjsdfasd
```

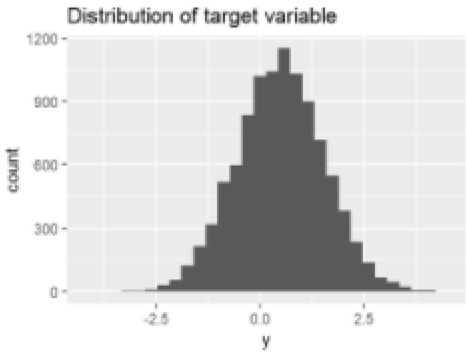
cannot find valid starting values: please specify some

Example: Gamma/Identity?

Q1 - Choose a distribution and link function.

Determine the best distribution and link function to use. Justify your choice of distribution based on the business problem and data and use only that combination for all further work. Test several combinations and select the one with the best AIC, QQ-plot, and graphs of residuals vs. fitted.

The distribution of the target variable Y is shown below, for $n = 10,000$ observations.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.9828 -0.2085  0.4971  0.4904  1.1956  4.4772
```

Sandbox | Start Over | Hints | Run Code

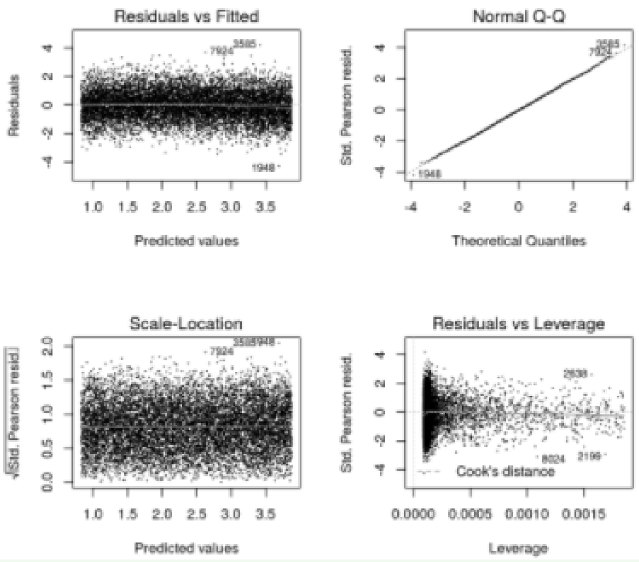
```
glm <- glm(formula=y ~ x, family = binomial(link = "identity"), data = glmdata1)
summary(glm)
par(mfrow = c(2,2))
plot(glm, cex= 0.1)
#lakjsdfasd
```

y values must be 0 <= y <= 1

Example: Gaussian/Inverse?

```
Call:
glm(formula = y ~ x, family = gaussian(link = "inverse"), data = glmdata1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.2538  -0.6879  -0.0159   0.6637   4.1848

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8740     0.1406   27.55   <2e-16 ***
x            -3.0431     0.1579  -19.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.02)

    Null deviance: 10981  on 9999  degrees of freedom
Residual deviance: 10195  on 9998  degrees of freedom
AIC: 28578

Number of Fisher Scoring iterations: 9
```

$AIC = 2k - 2\log like$



Interpretation Methods

1. Signs of coefficients
2. Sizes of coefficients
3. Probability modeling method (example cases)

Identity Link

For each one-unit increase in the variable X_j, the expected value of the target, E[Y], increases by β_j, assuming that all other variables are held constant.

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\frac{\partial E[Y]}{\partial X_1} = 0 + \beta_1 + 0$$

Log Link (Continuous X's)

Model Form

$$log(\hat{Y}) = X\beta \Rightarrow \hat{Y} = e^{X\beta}$$

$$g(\mu) = X\beta = \log(\mu) = \log(E(Y)) = \log(\hat{Y}) = X\beta$$

$$\mu = e^{X\beta}$$

Multiplicative Interpretation

$$\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}) =$$
$$e^{\beta_0} e^{\beta_1 X_{i1}} e^{\beta_2 X_{i2}} \ldots e^{\beta_p X_{ip}} = R_{i0} R_{i2} R_{i3} \ldots R_{ip}$$

Percentage Interpretation

Variable	β_j	$e^{\beta_j} - 1$	Interpretation
(intercept)	0.100	0.105	
X_1	0.400	0.492	49% increase in $E[Y]$ for each unit increase in X_1*
X_2	-0.500	-0.393	39% decrease in $E[Y]$ for each unit increase in X_2*

Log Link (Categorical X's)

	color=R	color=Blu
R	1	0
B	0	1
B	0	1
Y	0	0

Model Form

$$log(\hat{Y}) = X\beta \Rightarrow \hat{Y} = e^{X\beta}$$

$$\hat{Y} = 0.1 \; + 0.4\underbrace{(\text{Color} = RED)}_{X_1} \; - 0.5\underbrace{(\text{Color} = BLUE)}_{X_2}$$

Interpretation

Variable	β_j	$e^{\beta_j} - 1$	Interpretation
(intercept)	0.100	0.105	
Color=RED	0.400	0.492	49% increase in $E[Y]$ for RED cars as opposed to YELLOW cars*
Color=BLUE	-0.500	-0.393	39% decrease in $E[Y]$ for BLUE cars rather than YELLOW cars*

Example: June 16, 2020, Task 11

11. (*7 points*) Interpret the model for the client.

Run the recommended GLM from Task 7 on the full dataset.

- Copy the model output into your response.
- Interpret the coefficients for one categorical variable and one numeric variable to describe how these features relate to the target. The interpretations should be written in language appropriate for the client.

Example: June 16, 2020, Task 11

	β Estimate	e^{β} Exp(Estimate)
Coefficients:		
(Intercept)	0.80	2.22
x_1 genderMale	-0.03	0.97
x_2 age[10-20)	-0.10	0.91
x_3 age[20-30)	-0.23	0.79
x_4 age[30-40)	-0.25	0.78
age[40-50)	-0.23	0.79
x_5 readmitted<30	0.06	1.07
x_6 num_procs	0.01	1.01
x_7 num_meds	0.03	1.03
x_8 num_ip	0.01	1.01
x_9 num_diags	0.04	1.04

Interpretation
Start with a prediction of 2.2 days
If the patient is Male, multiply by 0.97
If the patient is age is between 10 and 20, multiply by 0.91
If the patient is age is between 20 and 30, multiply by 0.79
If the patient is age is between 30 and 40, multiply by 0.78
If the patient is age is between 40 and 50, multiply by 0.79
If the patient had been readmitted in the last 30 days, multiply by 1.07
For each procedure that the patient has had, multiply by 1.07
For each medication that the patient has had, multiply by 1.03
For each inpatient visit that the patient has had, multiply by 1.01
For each diagnosis that the patient has had, multiply by 1.04

$$\log(\mu) = \beta_0 + \beta_1(\text{gender}M) + \beta_2(\text{age}_{10-20}) + \cdots$$

$$\mu = e^{0.8}\underbrace{e^{-0.03x_1}}\, e^{-0.1x_2} \cdot \ldots \cdot e^{0.04(\text{num diags})}$$

GLMs for Classification

300-850 FICO® Score

The Score That Matters®

Your credit score: 850

Probability of Default: 1 - 850/1000 = 15%?

GLMs for Classification

β x xβ

Question	Variable	Coefficient	Sam's Value	Produc
How many credit cards do you have?	num_cards	4	5	20
How long ago did you get your first credit card?	length_of_credit	5	4	20
How long ago did you get your first loan? (i.e., auto loan, mortgage, student loan, etc.)	first_loan	8	5	40
How many loans or credit cards have you applied for in the last year?	num_loans	-0.01	2	-0.02
How recently have you opened a new loan or credit card?	new_cc	-0.005	1	-0.005
How many of your loans and/or credit cards currently have a balance?	num_cc_balance	-0.005	0	0
Besides any mortgage loans, what are your total balances on all other loans and credit cards combined?	total_balance	-0.01	5000	-50
When did you last miss a loan or credit card payment?	last_missed_pmt	-0.1	3	-0.3
What is the most delinquent you have ever been on a loan or credit card payment?	last_delinquent	-0.1	30	-3
How many of your loans and/or credit cards are currently past due?	num_past_due	-0.05	1	-0.05
What are your total balances on all currently past due accounts?	total_past_due	-0.1	300	-30
What percent of your total credit card limits do your credit card balances represent?	percent_balance	-0.05	5	-0.25
In the last 10 years, have you ever experienced bankruptcy, repossession or an account in collections?	bankruptcy	-0.2	0	0
				-3.625
	Linear Predictor (Z)	16.37		
	Probability of Default	0.5		$\boxed{2.60\%}$
	Sam's Credit Score			974
				750
			P(Not Default)	75%
			P(Default)	25%

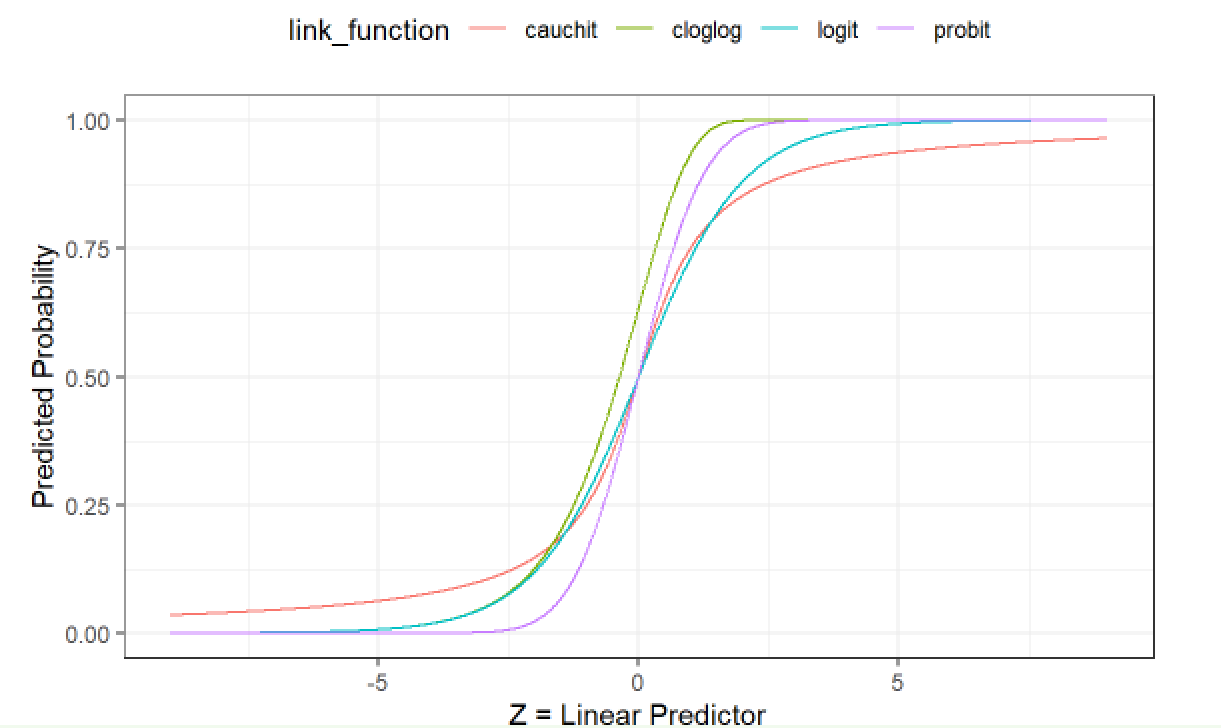
5(4) + 5(4) + ...

$$p = \frac{e^z}{1+e^z}$$

GLMs for Classification

$$E[Y] = p.$$

$$z = \log\left(\frac{p}{1-p}\right)$$



Quiz: Find the inverse

$$z = \log\left(\frac{p}{1-p}\right)$$

$$(1-p)e^z = \frac{p}{\cancel{(1-p)}}\cancel{(1-p)}$$

$$e^z - pe^z = p$$
$$+pe^z \qquad + pe^z \longrightarrow e^z = p + pe^z$$

$$e^z = p(1+e^z)$$

$$\Rightarrow \boxed{p = \frac{e^z}{1+e^z}}$$

Link Functions for Classification

Name	Link Function	Response Probability	Properties
Logit	$z = \log\left(\frac{p}{1-p}\right)$	$p = \frac{e^z}{1+e^z}$	Coefficients explained using odds; canonical link for binary family
Probit	$z = \Phi^{-1}(p)$	$p = \Phi(z)$	Coefficients explained as impact on z-score for Normal distribution
Cauchit	Na	$p = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$	Heavier tails than logit or probit
Cloglog	Na	$p = 1 - e^{-e^z}$	Inverse cdf of extreme value distribution; curv near probability of 1 is sharp

Other links (Probability Modeling / Example Cases Method)

Age	Years of Education	Internet Network	Hours Worked Per Week	Probability of High Profit
39	12	4G LTE	35	60%
53	18	4 G LTE	40	45%
25	16	5G	50	20%
27	16	5G	50	19%

Other links (Probability Modeling / Example Cases Method)

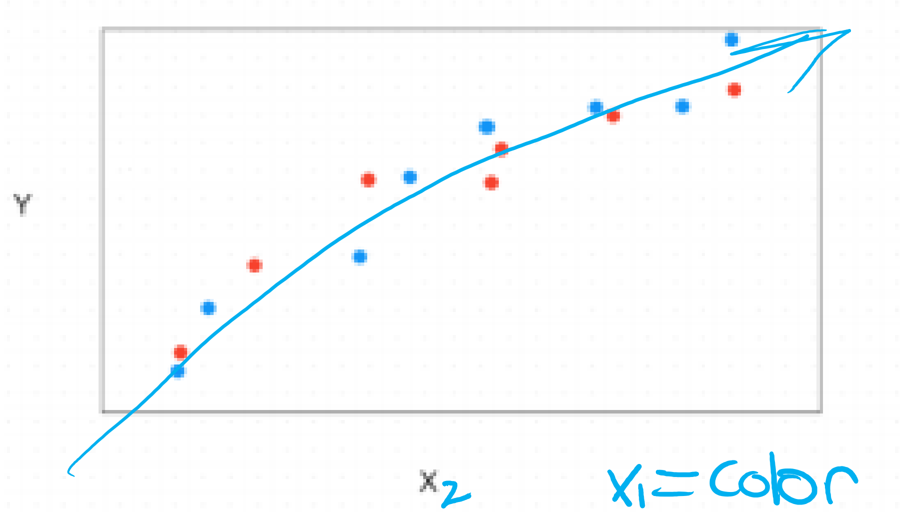
Age	Years of Education	Internet Network	Hours Worked Per Week	Probability of High Profit
39	12	4G LTE	35	60%
53	18	4 G LTE	40	45%
25	16	5G	50	20%
27	16	5G	50	19%
40	16	5G	50	30%

Interactions, Offsets, and Weights

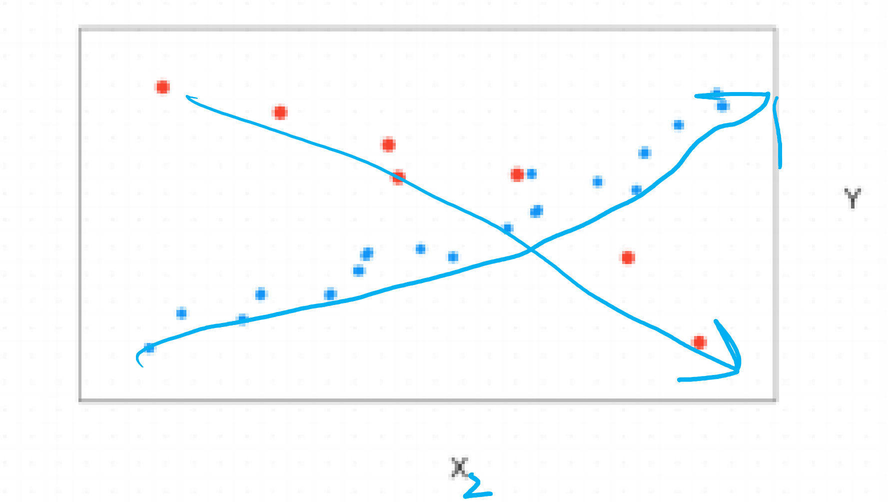
Offsets have *never appeared on exam pa*
Weights have *never appeared on pa*
Interactions appeared in June 2019, December 2019, June 2020

Interaction Terms

No interaction – slopes are the same



Interaction - Slopes are different



$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underline{\beta_3 x_1 x_2}$$

For **BLUES** (X_1 = 1):

$$\hat{y}_i = \beta_0 + \beta_1 + \beta_2 x_2 + \beta_3 x_2$$

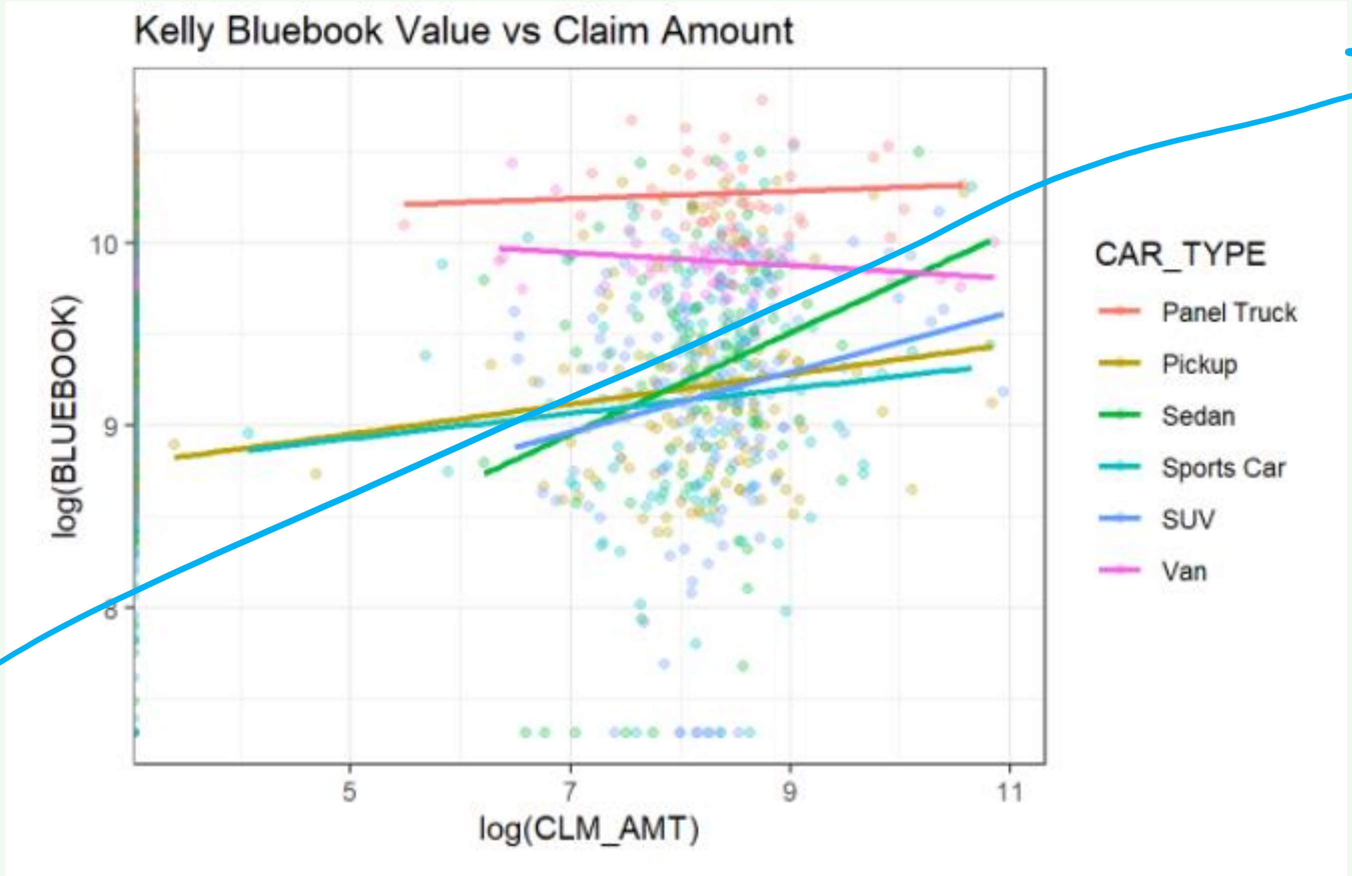
$$(\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_2$$

For **REDS** (X_1 = 0):

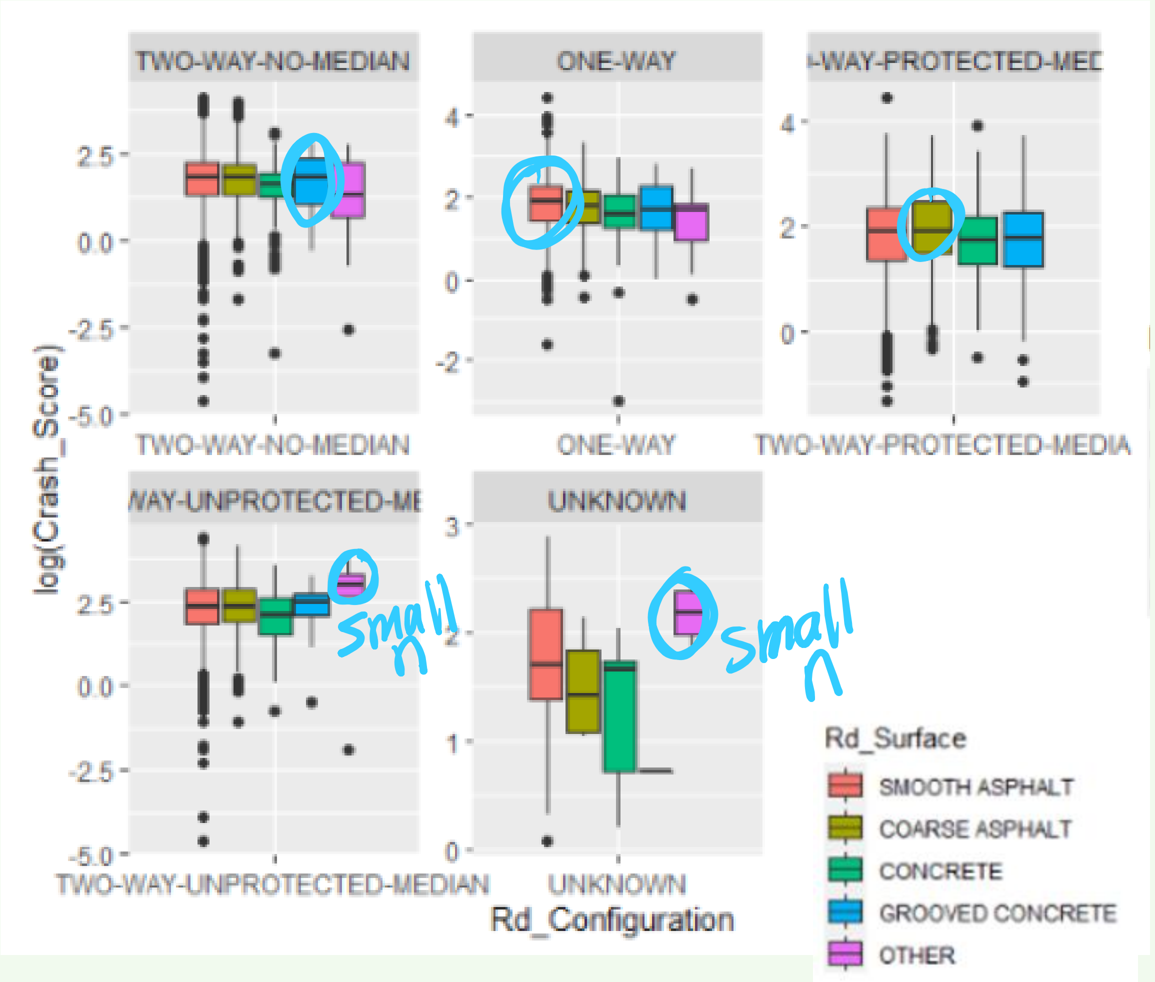
$$\hat{y}_i = \beta_0 + \beta_2 x_2 = \beta_0^* + \beta_1^* x_2$$

Interaction Example

Log(BLUEBOOK) ~ log(CLM_AMT)*CAR_TYPE



Interaction Example



Two-Way-No-Median: Highest crash score is for grooved concrete. This could be because the different road surfaces make turning more or less difficult. When there's no median on a two-way road, head-on collisions are possible. Grooved concrete may have less traction than smooth asphalt which makes these accidents more likely.
One-Way: Smooth Asphalt. The quality of the road will impact safety. Asphalt may be less expensive than concrete and so these roads are not as well maintained, have more potholes, and are not plowed as regularly following snowstorms.
Two-Way Protected Median: Coarse Asphalt. When there's a protected median, which I interpret as meaning that there is a guard rail dividing the two traffic directions, then head on collisions are not possible.
Two-Way Unprotected Median: Other (but there is a small sample size here so this may not be reliable).
Unknown: Also based on a small sample size.

Offsets

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \text{offset}(1)$$

Exam PA Assumptions:
#1: FAMILY = Poisson
#2: LINK = Log (Canonical)
#3: OFFSET = Exposure (usually length of policy period)

Y = # covid cases
W = # of People exposed to virus

$E\left[\frac{\text{\# covid cases}}{\text{\# of people exposed}}\right] = E\left[\frac{Y}{W}\right]$

$\log\left(\frac{E(Y)}{E(W)}\right) = X\beta$

$\log(E(Y)) - \log(E(W)) = X\beta$

$\log(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \boxed{\log(E(W))}$

Offsets example: Predict number of COVID cases

Y = Number of people infected with COVID
Exposure = Weights = Number of people exposed to the virus
Family = Poisson
Link = Log

Offsets vs. Weights

Both account for exposure (i.e., length of policy period, number of miles driven, number of insureds)

offset + loglink → log(exposure)

Weights:
Goal: Predict total claims (Severity).
Target = Average Claims (Claims Per-Member Per Month)
Weights = Member Months
Target*Weights = Total Claims (R does this automatically)

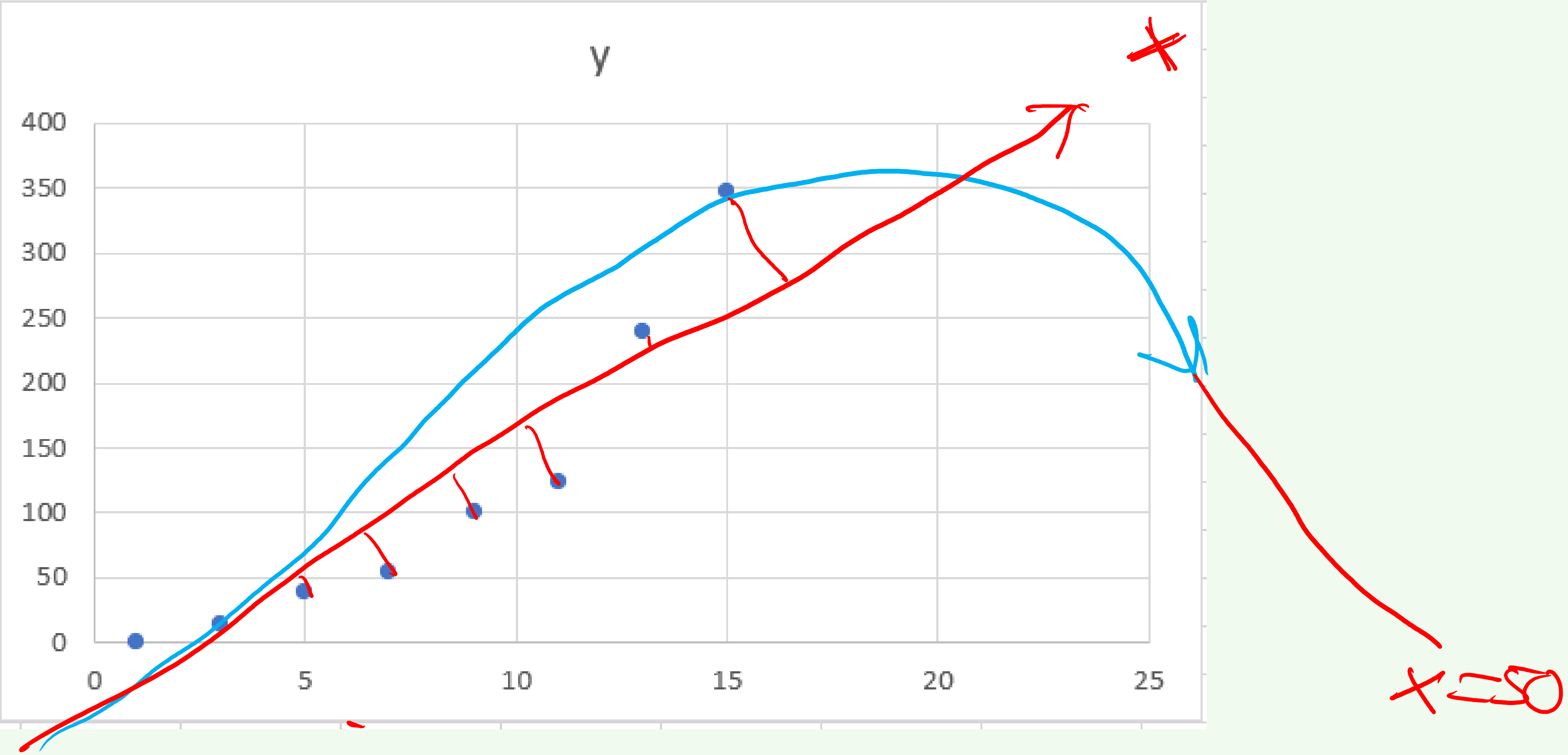
Offsets:
Goal: Predict number of claims (Frequency).
Target = Average Number of Claims (Claims Per-Member Per Month)
Offset = log(Member Months)

link = log

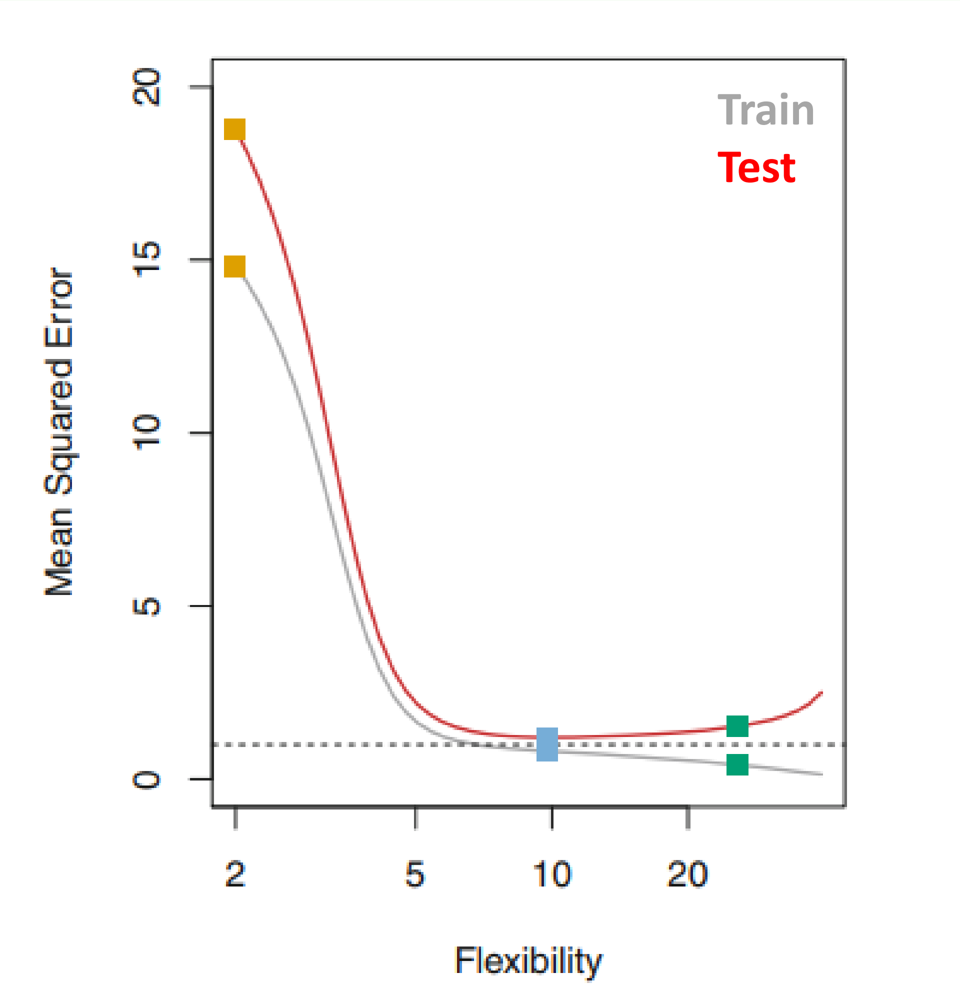
The Bias-Variance Trade-off

$R^2 \simeq 0.90$

$R^2 \simeq 0.8$



Training Error vs. Test Error



(a.k.a., “degrees of freedom”,
“number of predictors”, or “variance of model”)

Bias

"Not being complex enough to capture signal in the data"

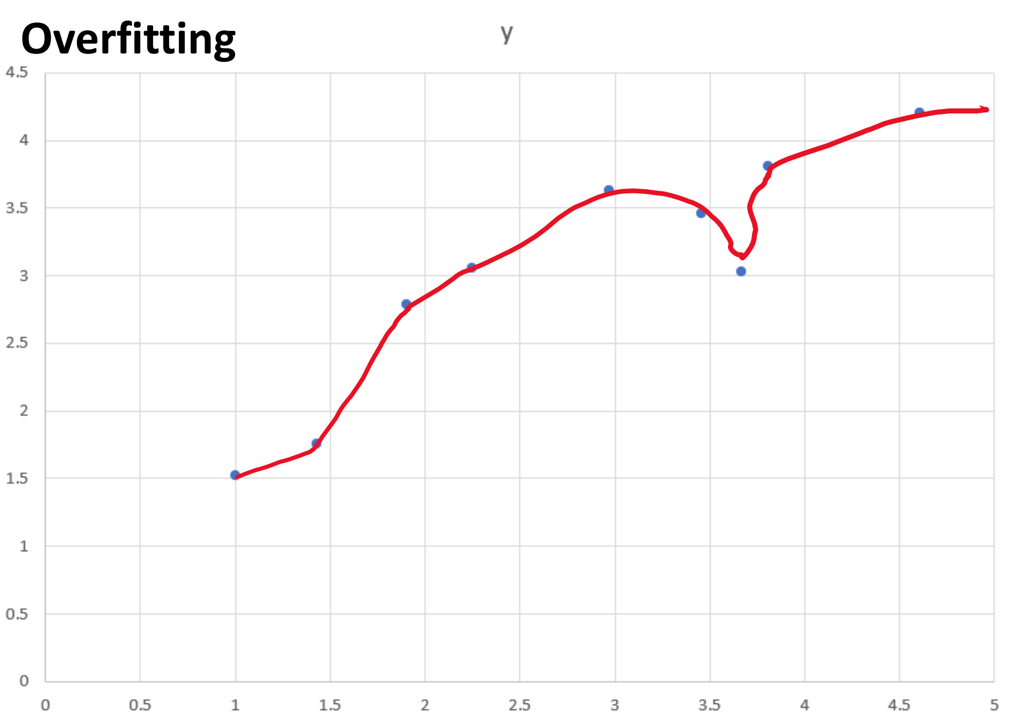
$Y = \text{Target}$

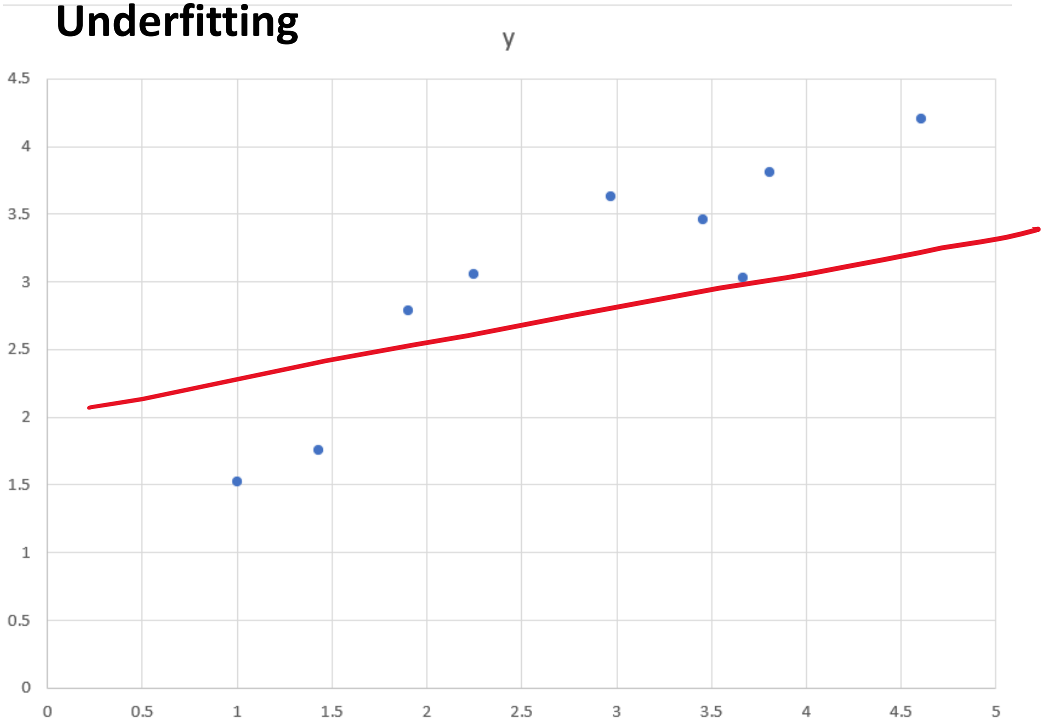
$f(X) = \text{Model}$

$\text{Bias} = E[Y] - E(f(X))$

Variance

"The expected loss from the model being too complex and overfitting to the training data"

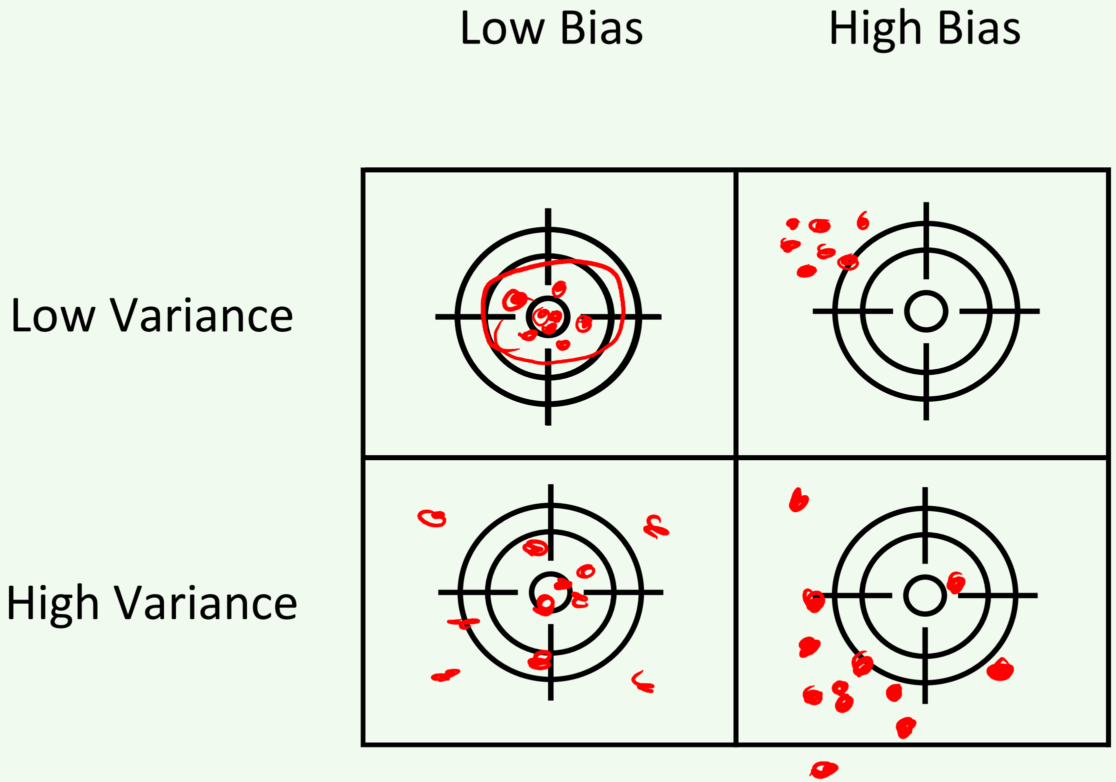




Bias-Variance Tradeoff

white noise

$$\text{Mean Squared Error} = \text{Variance of Model} + \text{Bias}^2 + \text{Irreducible Error}$$

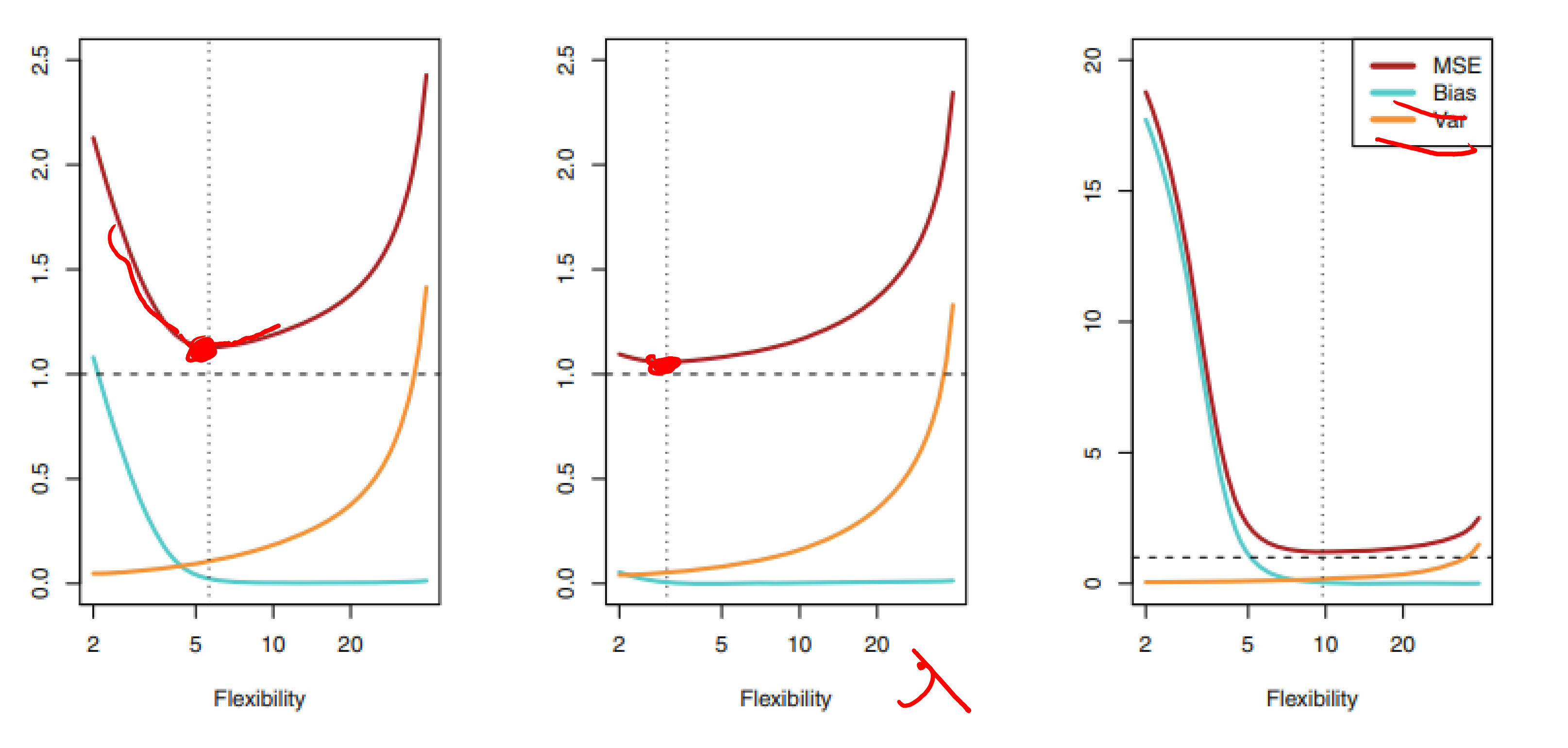


Bias-Variance Tradeoff

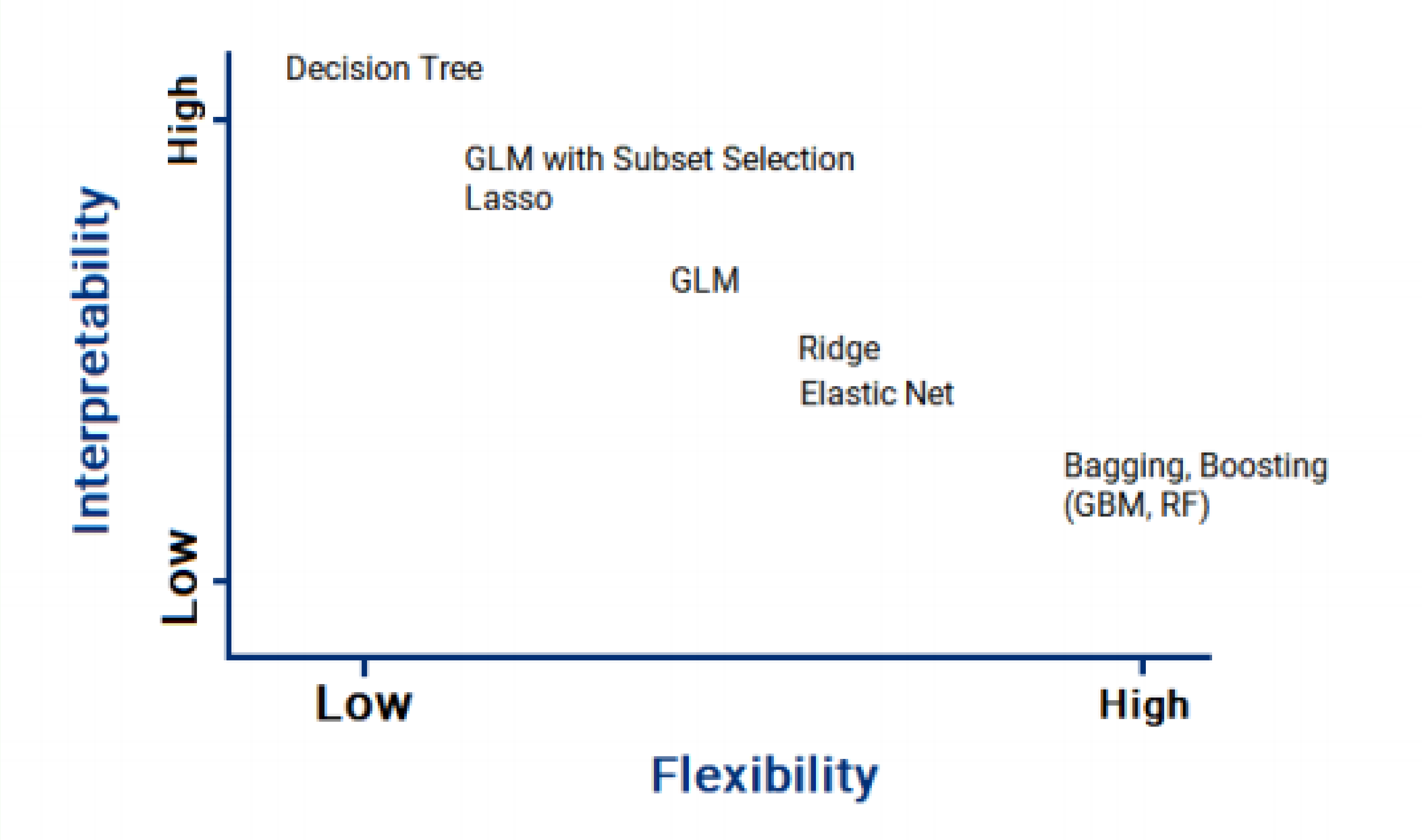
ISLR ch2

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

MSE



Bias-Variance Tradeoff



Example: June 16, 2020, Task 9 (7 points)

9. (*7 points*) Discuss the bias-variance tradeoff.

- Define bias and variance and describe the bias-variance tradeoff.
- Explain how lasso regression seeks to address the tradeoff.
- Explain how splitting the data into training and test sets and calculating a metric such as the Pearson goodness-of-fit statistic on the test set seeks to address the tradeoff.

Answer: June 16, 2020, Task 9 (7 points)

Bias is the expected loss caused by the model not being complex enough to capture the signal in the data. Variance is the expected loss from the model being too complex and overfitting to the training data.

We typically think of the expected loss as Bias + Variance + Unavoidable error. When building models, we are trying to minimize this expected loss, but to do so we often need to find a balance between bias and variance. Models with low bias tend to have higher variance and vice versa.

Without regularization, coefficients are found that maximize the likelihood function. This results in models that may not be optimal because coefficients are found even for features that may not be important. This process results in models that tend to overfit to the training data; they have high variance. LASSO penalizes models that have large coefficients to the extent that it can shrink coefficients of unhelpful predictors to zero. This is essentially trading some of the high variance from our non-regularized model for increased bias, which can potentially reduce the overall error.

With high variance (overfitting), the model will perform better on the training set than on a test set. With high bias (underfitting), the model will perform poorly on both the training set and the test set. When evaluating a single model, using a test set will help detect whether we have high variance because we can see a difference between the training and test set performance. When comparing models with different levels of complexity, comparing the test set performance and selecting the best performing model can also help us select the model design with the least total error.

Select Hyper-Parameters for Regularization Regression

- Lasso/Ridge/Elastic Net
- Step AIC

No memorization needed! Just use ?function_name in R

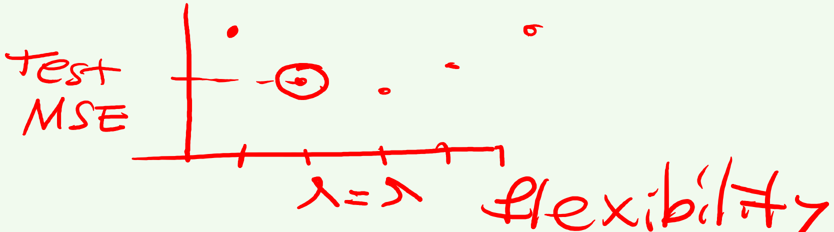
i.e.,
Library(glmnet)
?glmnet
Library(MASS)
?stepAIC
?AIC
? stepAIC

Shrinkage Methods

$$\text{RSS} = \sum_i (y_i - \hat{y})^2 = \sum_i \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2$$

Minimize MSE

No Shrinkage



RSS + λL2

$$\sum_i \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Ridge Regression (Make coefficients smaller)

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
$+ \beta_4 X_4$

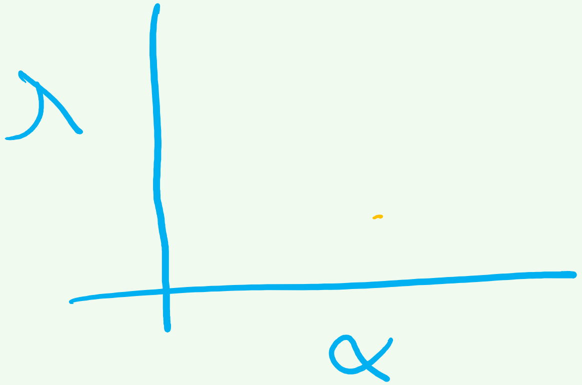
RSS + λL1

$$\sum_i \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso Regression (Make coefficients smaller and <u>exactly zero</u>)

Elastic Net (Best of both worlds)

$$\mathrm{RSS} + (1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|$$



Quiz: Hitters (See Study Guide)

1) How do ridge regression and the lasso improve on simple least squares?

2) In what cases would you expect ridge regression outperform the lasso, and vice versa?

Data #1

$n = 10{,}000$

$p = 500$

Data #2

$n = 5{,}000$

$p = 5$

Data #3

$P_{continuous} = 100$

$P_{Factor} = 2$

Data #4

$P_{continuous} = 1$

$P_{Factor} = 100$